

University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

**A Thesis Submitted for the Degree of PhD at the University of Warwick**

<http://go.warwick.ac.uk/wrap/36417>

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.

**COMPARING  
STATISTICAL METHODS AND ARTIFICIAL NEURAL NETWORKS  
IN BANKRUPTCY PREDICTION**

**BY**

**CHU, JUNG**

**B.B.A., M.A., M.S.**

**A THESIS SUBMITTED TO IN  
FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE  
OF**

**DOCTOR OF PHILOSOPHY**

**WARWICK BUSINESS SCHOOL**

**UNIVERSITY OF WARWICK**

**1997**

<b>TABLES OF CONTENTS</b>	
	<b>Page</b>
<b>LIST OF TABLES</b>	ix
<b>LIST OF FIGURES</b>	xvi
<b>ACKNOWLEDGEMENTS</b>	xxiv
<b>THESIS ABSTRACT</b>	xxv
 <b>CHAPTER 1 OVERVIEW</b>	
1.1 Introduction	1
1.2 Emerging Trends in Bankruptcy Prediction Techniques	2
1.3 The Objective of this Study	6
1.3.1 The Simulation Study	7
1.3.2 The Empirical Study	8
1.4 The Research Framework and the Key Research Questions	10
1.5 Organisation of the Thesis Chapters	15
 <b>CHAPTER 2 CONVENTIONAL STATISTICAL METHODS IN BANKRUPTCY PREDICTION</b>	
2.1 Linear Discriminant Function	17
2.2 Problems with the Discriminant Analysis	20
2.3 Studies Evaluating the Performance of MDA	21
2.3.1 Normality Assumption	21
2.3.2 Equal Group Dispersion Assumption	24
2.4 Nonparametric Discriminant Functions	27
2.5 The Reason for Selecting LDF as Discriminant Analysis in Our Study	28
2.6 The Importance of Prior Probabilities and Misclassification Costs	29

2.7 Logit Method	31
2.7.1 Logistic Regression	31
2.7.2 Other Link Functions and Corresponding Distributions	36
2.8 Comparisons of MDA and Logit	37
2.8.1 Theoretical Comparisons	37
2.8.2 Empirical Comparisons	39
2.9 Summary and Conclusions of MDA and Logit	43
 <b>CHAPTER 3 THE ARTIFICIAL NEURAL NETWORKS</b>	
3.1 Introduction of Artificial Neural Networks	44
3.2 The Basic Structure of A Neural Network	46
3.2.1 Layer and Node	46
3.2.2 Connection and Weight	47
3.2.3 Processing Element, PE	47
3.2.4 Network Operation	48
3.2.5 Feedforward Network vs. Recurrent Network	49
3.3 The Difference between ANNs and ES	49
3.4 Backpropagation and Generalised Delta Rule (GDR)	51
3.5 Variations on Standard Algorithm	52
3.5.1 Learning Rate	53
3.5.2 Momentum	53
3.5.3 Cumulative Update of Weight	54
3.5.4 Alternative Error Function	54
3.5.5 Different Activation Function	55
3.6 Problems with GDR Backpropagation	56



3.6.1 The Drawback on Long Learning Time	56
3.6.2 The Drawback on Local Minima Solutions	57
3.7 The Projection Neural Network	58
3.8 The Importance of Optimal Network Architecture	63
3.9 Past Studies on Determining Optimal Architecture	64
3.9.1 Trial and Error Approach	64
3.9.2 Genetic Algorithm	65
3.9.3 Weight Decay Technique	66
3.9.4 Pruning Technique	67
3.9.5 Principal Component Analysis (PCA)	68
3.9.6 Cascade Correlation Approach	70
3.10 The Significance Estimation of Input Variables	70
3.11 Incorporating Prior Probabilities and Misclassification Costs	72
3.12 Building A Backpropagation Neural Network	73
3.12.1 Network Design	74
3.12.2 Network Training and Optimal Structure Selection	74
3.12.3 Network Validation	76
3.12.4 Network Prediction	76
3.13 Summary and Conclusions	77
 <b>CHAPTER 4 COMPARISONS OF CONVENTIONAL STATISTICAL METHODS AND ARTIFICIAL NEURAL NETWORK</b>	
4.1 Introduction	79
4.2 Theoretical Comparisons of ANNs and STMs	81
4.2.1 Model Formulation and Problem Solving Procedure	81
4.2.2 Nonlinearity vs. Linear Boundary Building	89

4.2.3 Statistical Testing and Interpretation	94
4.2.4 Model Generalisation	96
4.2.5 Adaptability	97
4.3 Empirical Comparisons of ANNs and STMs	97
4.4 Summary and Conclusions	102

## **CHAPTER 5 MULTICOLLINEARITY AND FACTOR ANALYSIS**

5.1 Introduction	104
5.2 Multicollinearity in Bankruptcy	104
5.3 Factor Analysis	105
5.4 The Dilemma in Use of Factor Analysis	109
5.5 The Trade-off between Multicollinearity and Factor Analysis	112
5.6 Summary and Conclusions	113

## **CHAPTER 6 METHODOLOGY OF THE SIMULATION STUDY**

6.1 Motivation of The Simulation Study	118
6.2 Experimental Design	120
6.3 Data Set Generating	125
6.3.1 Data Distribution	125
6.3.2 Group Dispersion	128
6.3.3 The Relative Orientation between Predictor Variables	130
6.4 Parameter Selection and Correct Comparison	131
6.5 Techniques Description	133
6.6 Cutoff Point Determination	139

6.7 Statistical Test of Results	139
6.7.1 MANOVA Assumption	140
6.7.2 Effect Analysis	141
6.8 Summary and Conclusion	142

## **CHAPTER 7 RESULTS OF THE SIMULATION STUDY**

7.1 Introduction	150
7.2 The Relevant Hypotheses in Simulation Study	150
7.3 The Results and Analyses of Training Samples	151
7.3.1 Descriptive Statistics and Graphic Analyses	151
7.3.2 Change in Type I Error and Type II Error Rates	154
7.3.3 Multivariate Analysis	155
7.3.4 Univariate Analyses	156
7.3.5 Interaction Effects and Main Effects	157
7.4 The Results and Analyses of Testing Samples	177
7.4.1 Descriptive Statistics and Graphic Analyses	177
7.4.2 Change in Type I Error and Type II Error Rates	179
7.4.3 Multivariate Analysis	180
7.4.4 Univariate Analyses	180
7.4.5 Interaction Effects and Main Effects	180
7.5 Summary and Conclusions	196

## **CHAPTER 8 THE PROBLEMS OF BANKRUPTCY PREDICTION AND THEIR PROPOSED SOLUTIONS**

8.1 Introduction	198
8.2 Problems with the Selection of Predictor Variables	198

8.2.1 Accrual Accounting Ratios and Cash Flow Ratios	199
8.2.2 Theoretical Models of Bankruptcy	200
8.2.3 The Rationale for Choosing the Financial Ratios in this Study	205
8.3 Problems with Choice-Based Sampling Designs	206
8.3.1 Weighted Exogenous Sample Maximum Likelihood (WESML)	208
8.3.2 Proposed Solution to Determine Optimal Cutoff Point	211
8.4 Problems with Unequal Misclassification Costs	217
8.5 Problems with Model Validation and Generalisation	220
8.5.1 Two Validation Methods	221
8.5.2 <i>Ex post</i> Discrimination vs. <i>Ex ante</i> Prediction	222
8.5.3 The Impact of Different Base Rates between the Training and Testing Data Set	223
8.6 Summary and Conclusions	224

## **CHAPTER 9 METHODOLOGY OF THE EMPIRICAL STUDY**

9.1 Motivation of the Empirical Study	226
9.2 Data Collection	227
9.3 Descriptive Statistics of Data Set	227
9.3.1 Tests for Normality	229
9.3.2 Correlation Analyses and Variance-Covariance Matrices	234
9.3.3 The Difference in the Means of Twelve Financial Ratios between Failing and Nonfailing Firms	235
9.4 The Questions to be Tested and the Corresponding Research Designs	235
9.4.1 To Verify the Simulation Results in Real Financial Data Set	235
9.4.2 To Evaluate the Influence of Different Sample Size	236

9.4.3 To Investigate the Influence of Choice-Based Bias and WCOP Procedure	237
9.4.4 To Conduct a Sensitivity Analysis of Optimal Cutoff Points to Different Misclassification Costs	246
9.4.5 To Assess the Influence of Different Base Rate Between the Training and Validation Samples	247
9.5 Summary and Conclusions	248
 <b>CHAPTER 10 RESULTS OF EMPIRICAL STUDY</b>	
10.1 Introduction	250
10.2 The Results between ANNs and STMs Using Real Financial Data	250
10.2.1 Comparisons on Classification Accuracy	251
10.2.2 Comparisons on the Relative Importance of Predictor Variables	261
10.3 The Impact of Sample Size on Predictive Ability	267
10.3.1 The Results and Analyses	267
10.4 The Bias of Choice-Based Sample Design and Its Elimination by Applying WCOP Procedure	273
10.4.1 The Results and Analyses	274
10.5 The Sensitivity of Optimal Cutoff Points to Misclassification Costs	280
10.5.1 The Optimal Cutoff Points and Classification Accuracy for MDA Model	280
10.5.2 The Optimal Cutoff Points and Classification Accuracy for Logit Model	282
10.5.3 The Optimal Cutoff Points and Classification Accuracy for GDR Model	283
10.5.4 The Optimal Cutoff Points and Classification Accuracy for Proj Model	284
10.5.5 The Comparison of Type I Error to Different Relative Misclassification Costs among the Four Models	289

10.6 The Influence of Different Base Rate between Training and Testing Data on Classification Accuracy among the Four Methods	290
10.6.1 The Results and Analyses	291
10.7 Summary and Conclusions	305
<b>CHAPTER 11 OVERALL COMMENTS AND DIRECTIONS FOR FUTURE RESEARCH</b>	
11.1 Introduction	307
11.2 The New Elements in This Thesis	307
11.3 The Findings, Explanations and Implications	308
11.4 Data Conditions and Recommended Discriminating Techniques	319
11.5 Limitations of the Study	319
11.6 Directions for Future Research	321
<b>BIBLIOGRAPHY</b>	321
<b>APPENDIX</b>	
I. The Mathematical Derivation of BP Algorithm	347
II. Input Data for Example 1	350
III. Input Data for Example 2	351
IV. The Mathematical Derivation of an Alternative Projection.	353
V. The Z Scores for Each of 264 Companies and Descriptive Statistics by Group Using Multivariate Discriminant Analysis (MDA)	354
VI. The Intermediate Values Z of 264 Companies and Descriptive Statistics by Group Using Logit Procedure	356
VII. Conditional Probabilities of Bankruptcy for 264 Companies and Descriptive Statistics by Group Using Logit Procedure	359
VIII. The Intermediate Values Z of 264 Companies and Descriptive Statistics by Group Using GDR Neural Network Approach	362
IX. The Predicted Values vs. Actual Values of 264 Companies and Descriptive Statistics by Group Using GDR Neural Network Approach	365
X. The Intermediate Values Z of 264 Companies and Descriptive Statistics by Group Using Proj Neural Network Approach	368
XI. The Predicted Values vs. Actual Values of 264 Companies and Descriptive Statistics by Group Using Proj Neural Network Approach	371

## LIST OF TABLES

Tables	Page
2.6.1 Altman Cross-Validation Results	29
4.3.1 Summary of Previous Comparative Studies of ANNs and STMs	99
5.3.1 Summary of Representative Studies Using Factor Analysis to Extract Key Predictors for Evaluating the Financial Performance	110
5.4.1 Factor Analysis for Example 2 Data	117
6.4.1 Experimental Design of Simulation Study	131
6.4.2 Parameters of Experiment Data	143
7.3.1 Misclassification Summary in Normal Distribution Data Using Training Sample Results	165
7.3.2 Misclassification Summary in Skewed Distribution Data Using Training Sample Results	165
7.3.3 Misclassification Summary in Symmetric Distribution Data with Outliers Using Training Sample Results	165
7.3.4 The Pattern Of Type I and Type II Error for Each of Four Methods	166
7.3.I The Change and Percentage Change in Overall Error Rates between GDR, Proj and STMs Using Training Samples	166
7.3.II The Change and Percentage Change in Type I and Type II Error Rates between GDR and STMs Using Training Samples	166
7.3.III The Change and Percentage Change in Type I and Type II Error Rates between Proj and STMs Using Training Samples	166
7.3.5 Multivariate Analysis of Variance of MDA, Logit, GDR and Proj Methods on Overall Error Rates Using Training Data	167
7.3.6 Univariate Analysis of Variance for MDA Using Training Samples	167
7.3.7 Univariate Analysis of Variance for Logit Using Training Samples	167
7.3.8 Univariate Analysis of Variance for GDR Using Training Samples	167
7.3.9 Univariate Analysis of Variance for Proj Using Training Samples	167

7.3.10 The Three-Way Interaction Effects Dist by Disp by Orien for MDA Method Using Training Samples	171
7.3.11 Multiple Comparison for MDA Dist by Disp Interaction Effect Using Training Data	175
7.3.12 Multiple Comparison for MDA Dist by Orien Interaction Effect Using Training Data	175
7.3.13 Multiple Comparison for MDA Disp by Orien Interaction Effect Using Training Data	175
7.3.14 Main Effects for MDA Using Training Samples	175
7.3.15 Mean Error Rates for MDA Using Training Samples	174
7.3.16 The Three-Way Interaction Effects Dist by Disp by Orien for Logit Method Using Training Samples	172
7.3.17 Multiple Comparison for Logit Dist by Disp Interaction Effect Using Training Data	175
7.3.18 Multiple Comparison for Logit Dist by Orien Interaction Effect Using Training Data	175
7.3.19 Multiple Comparison for Logit Disp by Orien Interaction Effect Using Training Data	175
7.3.20 Main Effects for Logit Using Training Samples	175
7.3.21 Mean Error Rates for Logit Using Training Samples	175
7.3.22 Multiple Comparison for GDR Dist by Disp Interaction Effect Using Training Data	176
7.3.23 Multiple Comparison for GDR Dist by Orien Interaction Effect Using Training Data	176
7.3.24 Multiple Comparison for GDR Disp by Orien Interaction Effect Using Training Data	176
7.3.25 Main Effects for GDR Using Training Samples	176
7.3.26 Mean Error Rates for GDR Using Training Samples	176
7.3.27 Multiple Comparison for Proj Dist by Disp Interaction Effect Using Training Data	176



7.3.28 Multiple Comparison for Proj Dist by Orien Interaction Effect Using Training Data	176
7.3.29 Multiple Comparison for Proj Disp by Orien Interaction Effect Using Training Data	176
7.3.30 Main Effects for Proj Using Training Samples	176
7.3.31 Mean Error Rates for Proj Using Training Samples	176
7.4.1 Misclassification Summary in Normal Distribution Data Using Testing Sample Results	186
7.4.2 Misclassification Summary in Skewed Distribution Data Using Testing Sample Results	186
7.4.3 Misclassification Summary in Symmetric Distribution Data with Outliers Using Testing Sample Results	186
7.4.4 The Pattern Of Type I and Type II Error for Each of Four Methods	187
7.4.I The Change and Percentage Change in Overall Error Rates between GDR, Proj and STMs Using Testing Samples	187
7.4.II The Change and Percentage Change in Type I and Type II Error Rates between GDR and STMs Using Testing Samples	187
7.4.III The Change and Percentage Change in Type I and Type II Error Rates between Proj and STMs Using Testing Samples	187
7.4.5 Multivariate Analysis of Variance of MDA, Logit, GDR and Proj Methods on Overall Error Rates Using Testing Data	188
7.4.6 Univariate Analysis of Variance for MDA Using Testing Samples	188
7.4.7 Univariate Analysis of Variance for Logit Using Testing Samples	188
7.4.8 Univariate Analysis of Variance for GDR Using Testing Samples	188
7.4.9 Univariate Analysis of Variance for Proj Using Testing Samples	188
7.4.10 Multiple Comparison for MDA Dist by Disp Interaction Effect Using Testing Data	188
7.4.11 Multiple Comparison for MDA Dist by Orien Interaction Effect Using Testing Data	194
7.4.12 Multiple Comparison for MDA Disp by Orien Interaction Effect Using Testing Data	194

7.4.13 Main Effects for MDA Using Testing Samples	194
7.4.14 Mean Error Rates for MDA Using Testing Samples	194
7.4.15 Multiple Comparison for Logit Dist by Disp Interaction Effect Using Testing Data	194
7.4.16 Multiple Comparison for Logit Dist by Orien Interaction Effect Using Testing Data	194
7.4.17 Multiple Comparison for Logit Disp by Orien Interaction Effect Using Testing Data	194
7.4.18 Main Effects for Logit Using Testing Samples	194
7.4.19 Mean Error Rates for Logit Using Testing Samples	194
7.4.20 Multiple Comparison for GDR Dist by Disp Interaction Effect Using Testing Data	195
7.4.21 Multiple Comparison for GDR Dist by Orien Interaction Effect Using Testing Data	195
7.4.22 Multiple Comparison for GDR Disp by Orien Interaction Effect Using Testing Data	195
7.4.23 Main Effects for GDR Using Testing Samples	195
7.4.24 Mean Error Rates for GDR Using Testing Samples	195
7.4.25 Multiple Comparison for Proj Dist by Disp Interaction Effect Using Testing Data	195
7.4.26 Multiple Comparison for Proj Dist by Orien Interaction Effect Using Testing Data	195
7.4.27 Multiple Comparison for Proj Disp by Orien Interaction Effect Using Testing Data	195
7.4.28 Main Effects for Proj Using Testing Samples	195
7.4.29 Mean Error Rates for Proj Using Testing Samples	195
8.2.1 The Grouping of Financial Ratios Selected in Empirical Study	206
9.3.1 Descriptive Statistics for Each of 12 Financial Ratios	231

9.3.2 The Correlation Analysis between 12 Financial Ratios	232
9.3.3 The Variance-Covariance Matrices for Bankrupt and Nonbankrupt Firms	233
9.3.4 The Mann-Whitney U Test for Differences in the Means of 12 Financial Ratios Between Failing and Nonfailing Firms	234
9.4.1 The Research Design of Weighted and Unweighted Procedure on Six Choice-Based Estimation Samples	239
9.4.2 The Optimal Cutoff Points in MDA Approach	241
9.4.3 The Various K values in Logit Approach	243
9.4.4 The Optimal Cutoff Points in Logit Approach	243
9.4.5 The Optimal Cutoff Points in GDR Neural Network	245
9.4.6 The Optimal Cutoff Points in Proj Neural Network	245
9.4.7 The Research Design of Investigating the Influence of Different Base Rate between Training and Validation Samples on Predictive Ability	258
10.2.1 Comparison In Classification Performances Of Four Alternative Techniques for Training Sample and Testing Sample	251
10.2.2 The Frequency Distribution and Cumulative Distribution of Z Scores for MDA Method in Training Data	256
10.2.3 The Frequency Distribution and Cumulative Distribution of Z Scores for MDA Method in Testing Data	256
10.2.4 The Frequency Distribution and Cumulative Distribution of Conditional Probabilities for Logit Method in Training Data	256
10.2.5 The Frequency Distribution and Cumulative Distribution of Conditional Probabilities for Logit Method in Testing Data	256
10.2.6 The Frequency Distribution and Cumulative Distribution of Predicted Values for GDR Method in Training Data	257
10.2.7 The Frequency Distribution and Cumulative Distribution of Predicted Values for GDR Method in Testing Data	257
10.2.8 The Frequency Distribution and Cumulative Distribution of Predicted Values for Proj Method in Training Data	257

10.2.9 The Frequency Distribution and Cumulative Distribution of Predicted Values for Proj Method in Testing Data	257
10.2.10 The Pooled Variance-Covariance Matrix in MDA	263
10.2.11 Relative Contribution Tests of Each Independent Variables and Its Rank for MDA Method	264
10.2.12 Test of Individual Coefficients and Model Fitting for Logit Approach	265
10.2.13 The Neural Network Weights for GDR Method	265
10.2.14 The Neural Network Weights for Proj Method	265
10.2.15 Relative Strengths between Each Input Variables and Output for Both GDR and Proj Methods and the Rank of the Input Variables for Four Methods	266
10.2.16 Spearman Correlation Analysis for Ranks of Relative Importance of Predictor Variables among Four Methods	266
10.3.1 Misclassification Error Rates vs. Sample Size for Four Methods in Training Sample	268
10.3.2 Misclassification Error Rates vs. Sample Size for Four Methods in Training Sample	268
10.4.1 Comparison of Unweighted and WCOP on Classification Accuracy Across Alternative Estimation Samples in MDA Using Training Sample Results	276
10.4.2 Comparison of Unweighted and WCOP on Classification Accuracy Across Alternative Estimation Samples in MDA Using Testing Sample Results	276
10.4.3 Comparison of Unweighted and WCOP on Classification Accuracy Across Alternative Estimation Samples in Logit Using Training Sample Results	277
10.4.4 Comparison of Unweighted and WCOP on Classification Accuracy Across Alternative Estimation Samples in Logit Using Testing Sample Results	277
10.4.5 Comparison of Unweighted and WCOP on Classification Accuracy Across Alternative Estimation Samples in GDR Using Training Sample Results	278

10.4.6 Comparison of Unweighted and WCOP on Classification Accuracy Across Alternative Estimation Samples in GDR Using Testing Sample Results	278
10.4.7 Comparison of Unweighted and WCOP on Classification Accuracy Across Alternative Estimation Samples in Proj Using Training Sample Results	279
10.4.8 Comparison of Unweighted and WCOP on Classification Accuracy Across Alternative Estimation Samples in Proj Using Testing Sample Results	279
10.5.1 Summary of Optimal Cutoff Points and Accuracy for MDA Method	285
10.5.2 The Statistics of Z Scores Distribution for MDA Method Using 264 Companies	281
10.5.3 Summary of Optimal Cutoff Points and Accuracy for Logit Method	286
10.5.4 Summary of Optimal Cutoff Points and Accuracy for GDR Method	287
10.5.5 Summary of Optimal Cutoff Points and Accuracy for Proj Method	288
10.5.6 The Difference of Overall Accuracy between 1:1 Ratio and Critical Ratio for four Methods	289
10.6.1 Misclassification Rates of Different Base Rate between Training and Testing Data Composition Using Training Samples	292
10.6.2 Misclassification Rates of Different Base Rate between Training and Testing Data Composition Using Training Samples	293
10.6.3 The Difference between Type I and Type II Errors Over all Combinations	297
11.3.1 Summary of the Findings in this Thesis	309
11.4.1 Data Conditions and Recommended Discriminating Methods	320

## LIST OF FIGURES

Figures	Page
1.2.1 The Diagram of Three-layer Artificial Neural Network	4
1.4.1 The Framework of This Thesis	11
2.7.1 The Cumulative Distribution of Conditional Probability in Logit Procedure	33
2.7.2 The Hypothetical Logit Cumulative Density Function	36
3.1.1 The Perceptron Concept	45
3.7.1 The Projection Transformation and the Formation of Boundary Surface	59
3.7.2 The Composition of Training Process between GDR and Projection on the Circle and Rectangle Classification Problem	62
3.13.1 The Phases of Building A Neural Network	78
4.2.1 Simple Linear Perceptron = Multiple Linear Regression	86
4.2.2 Simple Nonlinear Perceptron = Logistic Regression	86
4.2.3 Perceptron with Threshold = Linear Discriminant Function	86
4.2.4 A Network with One Hidden Node	84
4.2.5 A Graph of $Y$ vs. $\ln(Y/(1-Y))$	85
4.2.6 Multilayer Perceptron = Simple Nonlinear Regression	87
4.2.7 Multilayer Perceptron = Multivariate Multiple Nonlinear Regression	87
4.2.8 Multilayer Perceptron = Nonlinear Regression Again	87
4.2.9 Multilayer Neural Network to Solve XOR Problem	92
4.2.10 The Decision Line in ANN to Solve XOR Problem	93
5.4.1 Scatter-Plots for Two Bivariate Normal Distribution	115
5.4.2 Univariate Distribution Plots Of $X_1$ for Two Groups	115
5.4.3 Univariate Distribution Plots of $X_2$ for Two Groups	115

5.4.4 Various 3-D Graphics for Example 2 Data	116
6.2.1 Three Orientation Schemes for Bivariate Variables Experimental Design	123
6.2.2 Experimental Design for Comparison of Classification Accuracy Between STMs and ANNs	124
6.4.1 Typical Aa1 Data Plot	144
6.4.2 Typical Aa2 Data Plot	144
6.4.3 Typical Aa3 Data Plot	144
6.4.4 Typical Ab1 Data Plot	144
6.4.5 Typical Ab2 Data Plot	144
6.4.6 Typical Ab3 Data Plot	144
6.4.7 Typical Ac1 Data Plot	145
6.4.8 Typical Ac2 Data Plot	145
6.4.9 Typical Ac3 Data Plot	145
6.4.10 Typical Ad1 Data Plot	145
6.4.11 Typical Ad2 Data Plot	145
6.4.12 Typical Ad3 Data Plot	145
6.4.13 Typical Ba1 Data Plot	146
6.4.14 Typical Ba2 Data Plot	146
6.4.15 Typical Ba3 Data Plot	146
6.4.16 Typical Bb1 Data Plot	146
6.4.17 Typical Bb2 Data Plot	146
6.4.18 Typical Bb3 Data Plot	146
6.4.19 Typical Bc1 Data Plot	147
6.4.20 Typical Bc2 Data Plot	147
6.4.21 Typical Bc3 Data Plot	147

6.4.22 Typical Bd1 Data Plot	147
6.4.23 Typical Bd2 Data Plot	147
6.4.24 Typical Bd3 Data Plot	147
6.4.25 Typical Ca1 Data Plot	148
6.4.26 Typical Ca1 Data Plot	148
6.4.27 Typical Ca1 Data Plot	148
6.4.28 Typical Cb1 Data Plot	148
6.4.29 Typical Cb2 Data Plot	148
6.4.30 Typical Cb3 Data Plot	148
6.4.31 Typical Cc1 Data Plot	149
6.4.32 Typical Cc2 Data Plot	149
6.4.33 Typical Cc3 Data Plot	149
6.4.34 Typical Cd1 Data Plot	149
6.4.35 Typical Cd2 Data Plot	149
6.4.36 Typical Cd3 Data Plot	149
6.5.1 A Diagram of the Optimal Structure of Neural Network in the Simulation Study	135
6.5.2 An Alternative Projection from 2-D onto 3-D	136
7.3.1 The Comparison of Four Methods for Normal Distribution on Training Data	168
7.3.2 The Comparison of Four Methods for Skewed Distribution on Training Data	168
7.3.3 The Comparison of Four Methods for Symmetric Distribution with Outliers on Training Data	168
7.3.4 The Comparison of Four Methods for (a) Group Dispersion on Training Data	168
7.3.5 The Comparison of Four Methods for (b) Group Dispersion on Training Data	168
7.3.6 The Comparison of Four Methods for (c) Group Dispersion on Training Data	168
7.3.7 The Comparison of Four Methods for (d) Group Dispersion on Training Data	169



7.3.8 The Comparison of Four Methods for Orientation I Scheme on Training Data	169
7.3.9 The Comparison of Four Methods for Orientation II Scheme on Training Data	169
7.3.10 The Comparison of Four Methods for Orientation III Scheme on Training Data	169
7.3.I The Plot of Percentage Change in Overall, Type I and Type II Error Rates between GDR and Statistical Methods Using Training Samples	170
7.3.II The Plot of Percentage Change in Overall, Type I and Type II Error Rates between GDR and Statistical Methods Using Training Samples	170
7.3.III No Interaction between A and B	157
7.3.IV Interaction between A and B	173
7.3.11 Interaction Effects of Dist by Disp on MDA Using Training Samples	173
7.3.12 Interaction Effects of Dist by Orien on MDA Using Training Samples	173
7.3.13 Interaction Effects of Disp by Orien on MDA Using Training Samples	173
7.3.14 Interaction Effects of Dist by Disp on Logit Using Training Samples	173
7.3.15 Interaction Effects of Dist by Orien on Logit Using Training Samples	173
7.3.16 Interaction Effects of Disp by Orien on Logit Using Training Samples	173
7.3.17 Interaction Effects of Dist by Disp on GDR Using Training Samples	174
7.3.18 Interaction Effects of Dist by Orien on GDR Using Training Samples	174
7.3.19 Interaction Effects of Disp by Orien on GDR Using Training Samples	174
7.3.20 Interaction Effects of Dist by Disp on Proj Using Training Samples	174
7.3.21 Interaction Effects of Dist by Orien on Proj Using Training Samples	174
7.3.22 Interaction Effects of Disp by Orien on Proj Using Training Samples	174
7.4.1 The Comparison of Four Methods for Normal Distribution on Testing Data	189
7.4.2 The Comparison of Four Methods for Skewed Distribution on Testing Data	189
7.4.3 The Comparison of Four Methods for Symmetric Distribution with Outliers on Testing Data	189

7.4.4 The Comparison of Four Methods for (a) Group Dispersion on Testing Data	189
7.4.5 The Comparison of Four Methods for (b) Group Dispersion on Testing Data	189
7.4.6 The Comparison of Four Methods for (c) Group Dispersion on Testing Data	189
7.4.7 The Comparison of Four Methods for (d) Group Dispersion on Testing Data	190
7.4.8 The Comparison of Four Methods for Orientation I Scheme on Testing Data	190
7.4.9 The Comparison of Four Methods for Orientation II Scheme on Testing Data	190
7.4.10 The Comparison of Four Methods for Orientation III Scheme on Testing Data	190
7.4.I The Plot of Percentage Change in Overall, Type I and Type II Error Rates between GDR and Statistical Methods Using Testing Samples	191
7.4.II The Plot of Percentage Change in Overall, Type I and Type II Error Rates between GDR and Statistical Methods Using Testing Samples	191
7.4.11 Interaction Effects of Dist by Disp on MDA Using Testing Samples	192
7.4.12 Interaction Effects of Dist by Orien on MDA Using Testing Samples	192
7.4.13 Interaction Effects of Disp by Orien on MDA Using Testing Samples	192
7.4.14 Interaction Effects of Dist by Disp on Logit Using Testing Samples	192
7.4.15 Interaction Effects of Dist by Orien on Logit Using Testing Samples	192
7.4.16 Interaction Effects of Disp by Orien on Logit Using Testing Samples	192
7.4.17 Interaction Effects of Dist by Disp on GDR Using Testing Samples	193
7.4.18 Interaction Effects of Dist by Orien on GDR Using Testing Samples	193
7.4.19 Interaction Effects of Disp by Orien on GDR Using Testing Samples	193
7.4.20 Interaction Effects of Dist by Disp on Proj Using Testing Samples	193
7.4.21 Interaction Effects of Dist by Orien on Proj Using Testing Samples	193
7.4.22 Interaction Effects of Disp by Orien on Proj Using Testing Samples	193
8.3.1 Hypothetical Frequency Distribution of Index I	214
8.4.1 Change of Optimal Cutoff Point to the Ratio of $C_{II}$ to $C_I$	219

10.2.1 Discriminant Function as an Equivalent Constraint Optimisation	253
10.2.2 Situation I of Dominant Discriminant Functions	254
10.2.3 Situation II of Dominant Discriminating Functions	255
10.2.4 Frequency Distribution of Z Scores for MDA Method in Training Data	258
10.2.5 Frequency Distribution of Conditional Probabilities for Logit Method in Training Data	258
10.2.6 Frequency Distribution of Predicted Values for GDR Method in Training Data	258
10.2.7 Frequency Distribution of Predicted Values for Proj Method in Training Data	258
10.2.8 Frequency Distribution of Z Scores for MDA Method in Testing Data	259
10.2.9 Frequency Distribution of Conditional Probabilities for Logit Method in Testing Data	259
10.2.10 Frequency Distribution of Predicted Values for GDR Method in Testing Data	259
10.2.11 Frequency Distribution of Predicted Values for Proj Method in Testing Data	259
10.2.12 Comparisons by Dominance for Training Data	260
10.2.13 Comparisons by Dominance for Testing Data	260
10.3.1 Error Rates vs. Sample Size in MDA for Training Data	270
10.3.2 Error Rates vs. Sample Size in MDA for Testing Data	270
10.3.3 Error Rates vs. Sample Size in Logit for Training Data	270
10.3.4 Error Rates vs. Sample Size in Logit for Testing Data	270
10.3.5 Error Rates vs. Sample Size in GDR for Training Data	270
10.3.6 Error Rates vs. Sample Size in GDR for Testing Data	270
10.3.7 Error Rates vs. Sample Size in Proj for Training Data	270
10.3.8 Error Rates vs. Sample Size in Proj for Testing Data	270
10.3.9 Type I Error Rates vs. Sample Size for four Methods for Training Data	271

10.3.10 Type II Error Rates vs. Sample Size for four Methods for Training Data	271
10.3.11 Overall Error Rates vs. Sample Size for four Methods for Training Data	271
10.3.12 Type I Error Rates vs. Sample Size for four Methods for Testing Data	271
10.3.13 Type II Error Rates vs. Sample Size for four Methods for Testing Data	271
10.3.14 Overall Error Rates vs. Sample Size for four Methods for Testing Data	271
10.5.1 The Graph of Z Scores in MDA	281
10.5.2 The Comparison of Sensitivity Analysis among Four Methods	290
10.6.I Overall Error for Various Base Rates in Training and Testing Data in MDA	301
10.6.II Overall Error for Various Base Rates in Training and Testing Data in Logit	301
10.6.III Overall Error for Various Base Rates in Training and Testing Data in GDR	301
10.6.IV Overall Error for Various Base Rates in Training and Testing Data in Proj	298
10.6.1 Type I Error vs. Various Base Rates of Testing Data when Base Rates of Training is Fixed to 1/1 Using Training Data Results	298
10.6.2 Type I Error vs. Various Base Rates of Testing Data when Base Rates of Training is Fixed to 1/5 Using Training Data Results	298
10.6.3 Type I Error vs. Various Base Rates of Testing Data when Base Rates of Training is Fixed to 1/9 Using Training Data Results	298
10.6.4 Type II Error vs. Various Base Rates of Testing Data when Base Rates of Training is Fixed to 1/1 Using Training Data Results	299
10.6.5 Type II Error vs. Various Base Rates of Testing Data when Base Rates of Training is Fixed to 1/5 Using Training Data Results	299
10.6.6 Type II Error vs. Various Base Rates of Testing Data when Base Rates of Training is Fixed to 1/9 Using Training Data Results	299
10.6.7 Overall Error vs. Various Base Rates of Testing Data when Base Rates of Training is Fixed to 1/1 Using Training Data Results	300
10.6.8 Overall Error vs. Various Base Rates of Testing Data when Base Rates of Training is Fixed to 1/5 Using Training Data Results	300
10.6.9 Overall Error vs. Various Base Rates of Testing Data when Base Rates of Training is Fixed to 1/9 Using Training Data Results	300

10.6.10 Overall Error for Various Base Rates in Training and Testing Data in MDA	301
10.6.11 Overall Error for Various Base Rates in Training and Testing Data in Logit	301
10.6.12 Overall Error for Various Base Rates in Training and Testing Data in GDR	301
10.6.13 Overall Error for Various Base Rates in Training and Testing Data in Proj	301
10.6.14 Type I Error vs. Various Base Rates of Testing Data when Base Rates of Training is Fixed to 1/1 Using Testing Data Results	302
10.6.15 Type I Error vs. Various Base Rates of Testing Data when Base Rates of Training is Fixed to 1/5 Using Testing Data Results	302
10.6.16 Type I Error vs. Various Base Rates of Testing Data when Base Rates of Training is Fixed to 1/9 Using Testing Data Results	302
10.6.17 Type II Error vs. Various Base Rates of Testing Data when Base Rates of Training is Fixed to 1/1 Using Testing Data Results	303
10.6.18 Type II Error vs. Various Base Rates of Testing Data when Base Rates of Training is Fixed to 1/5 Using Testing Data Results	303
10.6.19 Type II Error vs. Various Base Rates of Testing Data when Base Rates of Training is Fixed to 1/9 Using Testing Data Results	303
10.6.20 Overall Error vs. Various Base Rates of Testing Data when Base Rates of Training is Fixed to 1/1 Using Testing Data Results	304
10.6.21 Overall Error vs. Various Base Rates of Testing Data when Base Rates of Training is Fixed to 1/5 Using Testing Data Results	304
10.6.22 Overall Error vs. Various Base Rates of Testing Data when Base Rates of Training is Fixed to 1/9 Using Testing Data Results	304
11.3.1 95% Confidence Interval of Average Misclassification Rate for MDA, Logit, GDR and Pro Methods Using Training Data	313
11.3.2 95% Confidence Interval of Average Misclassification Rate for MDA, Logit, GDR and Pro Methods Using Testing Data	314
11.3.3 Discriminant Functions on Aa1 Data Sets	316
11.3.4 Discriminant Functions on Ab1Data Sets	316
11.3.5 Discriminant Functions on Ac Data Sets	316
11.3.6 Discriminant Functions on Ad Data Sets	316

## ACKNOWLEDGEMENTS

This research could not have been accomplished without the help and support of several individuals. First, I would like to express my deep gratitude to my major supervisor, Professor Anthony Steele, for his continuous guidance and encouragement. His penetrating comments and ample knowledge have greatly enhanced the quality of this study and my academic experience at the University of Warwick. His assistance will never be forgotten.

Special appreciation is expressed to my the other supervisor, Dr. Robert Hurron, for his advice, insights in the development of this thesis.

I would also like to thank to my examiners: Professor R. H. Berry and Dr. CGC Pitts. Their valuable suggestions and comments have made this thesis a great improvement.

This thesis is dedicated to my husband, Mao-Kuen Kuo, and our daughters, Nancy and Hsuan-Hsuan. Without their support, understanding and love, the completion of this study would not have been possible. I shall remember forever the sacrifices that they made so that my educational objectives can be attained.

Gratitude is also expressed to my parents for providing me with a loving home and supporting my early education.

Finally I thank God for giving me courage to undertake this PhD program and leading me all the way to the completion of this thesis.

## THESIS ABSTRACT

The use of multivariate discriminant analysis (MDA) and logistic regression procedure (Logit) in predicting business failure has been explored in numerous studies since 1960s. Recently, a newly developed technique, artificial neural networks (ANNs), has attracted much attention and has been applied to bankruptcy prediction area. At the same time, many papers attempted to compare the predictive ability of these two distinct classes of discriminators in order to find a best failure prediction method. However, most of their results, despite showing the superiority of ANNs, have been sharply criticised either for the unfair comparison or for their specific data selection. There is a need to undertake theory-based research to identify problem characteristics that predict when ANNs will forecast better than statistical models; to identify which input variable characteristics predict when ANNs will improve model estimation; and to identify when this advantage would give substantially improved forecasting performance.

Motivated by the limited amount of research on investigating the relative effectiveness of traditional methods as compared to the ANNs under a wide variety of modelling assumptions, one of the objectives of this study is to compare their classification capacities on a theoretical basis, and to evaluate the robustness on certain situations through the simulation study. The investigation is conducted on two popular statistical techniques—the MDA and the Logit, as well as two different learning algorithms of ANNs—the standard generalised delta rule (GDR) and the Projection approach (Proj). This can be regarded as the horizontal assessments of bankruptcy prediction.

The other aim of this thesis is to evaluate the impacts of variations in failure prediction models through the empirical study. These variations involve the issues we often encounter in the real world, such as the different sizes of sample, a choice-based sampling bias, the sensitivity of optimal cutoff points to misclassification costs of Type I and Type II errors, and the imbalance of the composition of failed to nonfailed firms between training and testing data sets. This can be viewed as the vertical assessments of bankruptcy prediction.

The simulation results indicate that the neural networks are indeed competitive approaches on bankruptcy prediction. In particular, the Projection network, which was developed to overcome the drawbacks that a commonly used GDR backpropagation algorithm often experiences, proves its remarkable superiority not only quantitatively (i.e., lower overall accuracy), but also qualitatively (lower Type I and Type II errors). The Projection network holds a promise for future elaboration.

Moreover, the outcomes of empirical experiments enhance our knowledge of some factors in constructing a failure forecasting model. This knowledge is related to both traditional statistical tools and modern neural networks and is essential for decision making.

# **Chapter One**

## **OVERVIEW**

### **1.1 Introduction**

This thesis is divided into two major parts. First, the classification performance of two statistical methods (STMs) and two artificial neural network (ANNs) will be evaluated and compared using comprehensive simulation data on bankruptcy prediction. Secondly, in order to validate the simulation results, an empirical study based on real financial data will be conducted. In addition, some variations in the construction of bankruptcy prediction models in practice will be discussed. Further, the problems resulting from these variations, which could affect the usefulness of a model's assessment, will be coped with by using our proposed solutions.

These investigations augment previous applications of multivariate discriminant analysis (MDA), Logit procedure and artificial intelligence techniques in accounting research by

1. evaluating whether the ANN is a promising solution in predicting business failure through generating simulated financial ratios which vary by (a) data distribution, (b) relationship of variance-covariance matrices across different groups, and (c) relative orientations between predictor variables;
2. investigating the impact of sample size on reliability of predictive capacity;
3. dealing with the choice-based sampling bias problem;
4. testing the sensitivity of optimal cutoff points to misclassification costs of Type I and Type II errors;
5. examining the generalisation capability on the condition that there exists an imbalance in the composition of failing to nonfailing firms between training and testing samples.

In this thesis we apply a newly developed methodology to an old accounting problem and assess simulation as a method for investigation in this area.



## **1.2 Emerging Trends in Bankruptcy Prediction Techniques**

Predictions of firm bankruptcy have been extensively studied in finance, accounting and decision sciences over the past two decades. Creditors, investors, auditors, regulators and managers need models to analyse the financial performance of firms. Creditors are concerned about the health of a borrower's financial situation in order to make a loan decision and to monitor performance during the period of repayment. Auditors can apply such a model in order to issue an auditing statement concerning going concern assumptions. In the United States auditing guidelines for assessing the going-concern status of business entities were established only in 1981 through Statement on Auditing Standards and are in SAS No.34. A bankruptcy prediction model can be suggested for aiding auditors in making going-concern assessments. Investors use the firm's distress prediction information to choose corporate debt and equity securities. Government regulators as well as managers benefit from an early warning provided by a bankruptcy prediction model.

In order to improve the models, previous researchers put a great deal of effort into empirical studies using traditional statistical techniques as well as newly developed neural network methods. Beaver [1966, 1968] and Altman [1968] established the foundation for subsequent methodological development in bankruptcy modelling. In Beaver's study [1968], a univariate financial variable was first used to measure the ability to predict business failure. Altman [1968] developed the multivariate model using multivariate discriminant analysis (MDA). Instead of employing univariate measurement, MDA attempts to derive a linear combination of multivariate characteristics of an observation and then classifies the observation into the appropriate prior defined group. It is based on the entire profile of characteristics and the interactions between the characteristics rather than on just an individual attribute as developed by Beaver [1968]. Subsequently the MDA was given wide application in studies by Edmister [1972], Deakin [1972, 1976], Blum [1974], Pinches et al. [1973, 1975], Bird and McHugh [1977], Altman et al. [1977] and Altman and Eisenbeis [1978]. However, Ohlson [1980] made a significant change in the statistical methods for evaluating business failure prediction models in his research. He used Logit procedure in place of the traditional MDA methodology. Logit procedure is a conditional probability approach which provides an estimate of the probability that an observation will

enter, given the values of the explanatory variables for the observation in question. It does not share the strict MDA assumptions of multivariate normality and homogeneity of variance-covariance matrices. Besides, it yields a meaningful interpretation of the significance of individual variable coefficients. As a consequence there has been considerable research employing the Logit approach, including Mensah [1983], Zavgren [1985], Casey and Bartczak [1985], Storey et al. [1987], Peel [1987], Zavgren et al. [1988] and Keasey and McGuinness [1990].

As an alternative to statistical approaches, artificial intelligence (AI) has been applied to accounting problems in much of the research. This has involved the creation of an expert system. An early expert system within the discipline of accounting was proposed and built by Hansen and Messier [1986] for the purpose of analysing EDP audit activities and assisting in auditor opinion formulation. The main task in developing an expert system has involved the "extraction" of expertise from professionals knowledgeable within the specified domain. The process of extracting knowledge from domain expertise is typically time-consuming and error-prone [Hayes-Roth et al., 1983]; [Greene, 1987]. Moreover, even domain experts sometimes have difficulty in establishing appropriate values in order to achieve a solution. This is called "the paradox of expertise" [Johnson, 1983].

Another form of AI, artificial neural network (ANN), has been newly developed and utilised. A neural network, inspired by biological neurone systems, consists of a set of elementary processing units which operate in a parallel distributed processing manner. Parallel processing appears to provide a powerful and practical approach which can overcome the speed limitation of a single processor. Further, unlike an expert system that requires explicit inference rules or mechanisms, a neural network has inherent learning ability to adapt behaviour in accordance with observations. Due to the features of this powerful learning ability, the capacity to "see through" noise and distortion [Wassermann, 1989], a strong adaptability, a distributed associative memory, and a high degree of robustness and fault tolerance [Lippman, 1987], the neural network has been used increasingly in the area of bankruptcy prediction [Odom and Sharda, 1989]; [Salchenberger et al., 1992]; [Bell et al., 1990]; [Tam and Kiang, 1992]; [Coats and Fant, 1993]; [Wilson and Sharda, 1994]; [Altman et al., 1994]; bond rating [Dutta and Shekhar, 1988]; [Surkan, and Singleton, 1989]; [Utans and Moody, 1991]; and in security price forecasting [Yoon et al., 1993]; [Refenes, Francis and Zapranis, 1994]; etc.

One of the most effective learning algorithms in computing with neural nets is backpropagation neural network (BPNN) [Parker, 1985]; [Rumelhart et al., 1986]. A standard generalised delta rule (GDR) backpropagation neural network is composed of at least three levels of units: an input layer, a hidden layer and an output layer. A simple diagram of a three-layer artificial neural network is presented in Figure 1.2.1.

Each input node is fully connected to each node in the hidden layer, and each unit in the hidden layer is connected to each unit in the output layer. The connection in the two nodes is called a link. A weight is attached to each link, which has the effect of attenuating or amplifying a transmitted value. It represents the stored knowledge of the model in the neural network. When the inputs flow through the network to the output layer, the differences between the computed output and desired output are calculated, and the weights are adjusted in a backward direction based on the gradient descent rule.

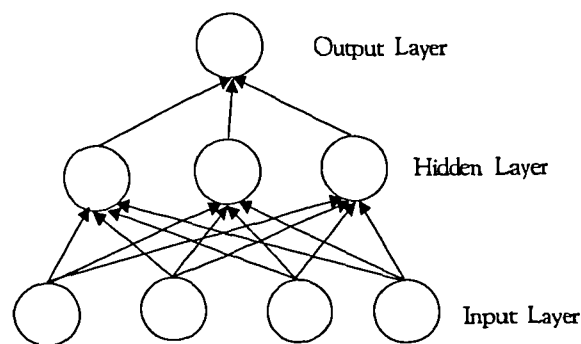


Figure 1.2.1 The Diagram of a Three-Layer Artificial Neural Network

The number of units in the input and output layers depends on the problem we want to solve. But the units in the hidden layer are chosen arbitrarily and are closely related to the network capacity. [Akaho and Amari, 1990]; [Baum and Wilczek, 1988]. Too many hidden units may result in a solution that cannot be generalised [Caudill, 1990]. On the other hand, too few hidden units may cause a larger number of training iterations and a very slow training process, or may even result in the network's inability to produce a desired output function. Therefore, an optimal architecture should be determined and built before using the neural network technique.

Several studies have proved that the three-layer network with a finite number of hidden nodes can represent any continuous function at any degree of desired accuracy [Cybenko, 1989]; [Hecht-Nielsen, 1989]; [Hornik et al., 1990]. It accomplishes this by partitioning the input space with hyperplanes, and the nonlinear combination of such hyperplanes can partition the input space into closed and open regions bounded by hyperplanes and curved surfaces. More specifically, the novelty about neural networks lies in their ability to model non-linear pattern with no prior assumptions about the nature of the generating process.

However, major problems are encountered with the backpropagation algorithm: namely, its long training time and the difficulty of getting trapped in a local minimum of the error function. The most important factor causing local minima or a slow learning speed is the inappropriate and aimless randomisation of the initial weights and thresholds. As a result, derivative-based search techniques commonly used in BPNN tend to become stuck at local minimum or take a long time to achieve a global solution. Lee et al. [1991] have indicated that the initial weight has a direct effect on the training speed and convergence to local minima. Kolen and Pollack [1990] have also demonstrated that backpropagation is very sensitive to the initial weights. In a GDR backpropagation neural network, initial weights and thresholds are usually chosen randomly. Thus, finding a method which can cause a good initial point is a sensible way to solve these two problems.

A different training algorithm developed by Wilensky and Manukian [1992] is called Projection neural network (Proj), which overcomes the drawbacks of BP-trained network. The main idea behind this network is to build a more sophisticated and faster network by combining the advantages of hypersphere classifiers (closed boundary networks), such as RCE (reduced Coulomb energy) [Reilly et al., 1982] and ART (adaptive resonance theory) [Carpenter and Grossberg, 1987], and backpropagation ANN (open boundary network) in a single network. The hypersphere classifiers have the advantage of training quickly because of their forming a closed decision boundary. In contrast the BPNN algorithm offers the advantage of ensuring error minimisation when it converges. The Projection algorithm can thus rapidly place the closed decision boundary prototype around the input points and minimise the output error through gradient descent. It initialises the weights and thresholds in higher dimension space as a prototype with a closed or open boundary, providing a good starting-point so that the network output is already close to a desired

output and avoids local minima which GDR backpropagation often experiences. In another words, the Projection algorithm is developed by projecting the input vector of a standard neural network onto a hypersphere in one higher dimension to form closed prototypes, and then backpropagation training is used to adjust the network weights and thresholds to ensure error minimisation. Wilensky and Manukian [1992] also described the attraction that this algorithm is able to combine closed prototypes with open prototypes, and thus requires only one node per closed region instead of the large number of nodes required in BPNN when the classification boundary is complex. From a theoretical point of view, it seems to provide a more effective approach to the classification problem by allowing a good initial setting of the weights and thresholds and time saving in the training sample [Wilensky and Manukian, 1992]. However, its classification accuracy has neither been empirically studied nor verified through comprehensive data.

### **1.3 The Objective of This Study**

Recently the comparison of conventional statistical methods such as MDA, Logit and artificial neural networks has stimulated much interest. To date, many papers in the areas of accounting and finance have contributed to this task. The intention of theses studies is to compare the predictive abilities of these two distinct classes of techniques and to find the best corporate failure prediction method. However, their results have been sharply criticised either for unfair comparison or for their specific data selection. That is, the comparisons of the predictive abilities of these methods have been validated only by their empirical performance and not on any theoretical basis. It is time to go beyond case studies and to undertake theory-based research in order to identify the advantages and disadvantages of these techniques; to identify when this advantage would give substantially improved forecasting performance; and to isolate those situations that certain technique is unable to produce effective classification ability.

One of the objectives of this study is to compare predictive ability through extensive data situations and to evaluate robustness on certain data conditions for MDA, Logit and two different learning algorithms of ANNs. The other aim is to investigate the impacts of

variations experienced in bankruptcy prediction models in real world. These variations involve the issues such as variable sample size, choice-based sampling bias, unequal Type I and Type II error costs, and an imbalance in the composition of failed to nonfailed firms between training and testing data. There is a need to understand these merits and shortcomings in order to avoid unfavourable circumstances and to benefit from strengths. These aims will be achieved through analysis of both a simulation and an empirical study.

### **1.3.1 The Simulation Study**

The simulation study, the first part of this research, will be conducted in two stages. In the first stage a broad range of data conditions will be generated based on the Monte Carlo resampling experiments. Bivariate-population data sets will be produced by employing three variables: the levels of data distribution, group dispersion and orientation between predictors. The rule which is applied to generate each of these controlled variables will be based on the survey in the previous study. In other words, data generation is justified by means of creating data in such way that as far as possible there is no departure from the real situation of the financial data.

The second stage of the experiment will test the classification accuracy of each of the aforementioned techniques using these varieties of simulation data. Three different measures of classification accuracy will be determined: Type I error (the error of misclassifying a failure company as a healthy company), Type II error (the error of misclassifying a healthy company as a failed company), and Overall error rate (the total error divided by total observation tested). They are provided by the specified procedure in terms of both training set and testing set, and used to compare the predictive capability for each of the four methods. Since the robustness test and the classification accuracy comparison form the primary emphasis, in order to simplify the simulation process, the experiments will assume equal misclassification cost and prior probability.

The simulation study involves a factorial design which is subjected to a three-way (three variables), fixed effects analysis of variance design with replicated observations per cell. To clearly display the main effect of all factors (variables) and their interactions, further analysis will be conducted using the Multivariate Analysis of Variance (MANOVA) statistical tool.

The goals of the simulation study are

- (1) To establish whether the ANN is a more competitive procedure under a broad range of modelling assumptions in bankruptcy prediction.
- (2) To assess whether the Projection algorithm holds a promise for future elaboration.
- (3) To identify the problem and input variable characteristics for which the ANNs and statistical methods are able or unable to produce effective discriminating capacity, and to suggest possible modifications.

### **1.3.2 The Empirical Study**

The primary aim of the simulation study is to test the robustness of alternative techniques and to compare their classification accuracy in bankruptcy prediction under comprehensive data situations. Since we can hardly collect such versatile data sets in the real world to conduct these tests, simulation is a good tool to achieve this goal. However, simulation has its limitations. The data sets generated through simulation are unlikely to cover the complicated covariance structure among indicators found in real data. Additionally, some impractical assumptions, such as the equal prior probability of two groups and the equal misclassification cost of two type errors need to be overcome and must be further explored.

Thus the empirical study, the second part of this thesis, will consolidate the simulation study. One of its intention is to validate the results from the simulation study. The other intention is to investigate the impact of some variations in the bankruptcy prediction model's classification accuracy for traditional as well as modern discriminating techniques.

The research will build on the data collected by Lin [1993], which was extracted from a UK business database. The financial indicators will be selected on the basis of bankruptcy theories which have proven successful performance in previous empirical studies.

The objectives of the empirical study are

- (1) To investigate the impact of sample size on predictive capabilities.

The number of observations needed to develop a reliable model is often one of the main concerns of practitioners. This concern is especially common for ANNs due to their requiring large sample sizes based on some prior research. There is no magic formula to determine what size will be appropriate for each different case. Most

studies choose their sample size in terms of the availability of data. Establishing a general rule is not our goal because it seems to be largely dependent upon the complexity of the problem being solved. However, an understanding of the effects of sample size on prediction performance is necessary and helpful.

- (2) To assess the influence of choice-based sampling bias and eliminate this bias using proposed adjustment procedure.

With most bankruptcy modelling, the proportion of nonfailing to failing firms in a sample (base rate) has nothing to do with the frequency of these two groups in an actual population. Different base rates are incorporated because of data availability. On the other hand, bankrupt firms are matched with the other group (nonbankrupt firms) on the basis of certain rules for controlling industry and size effect. Some presumed prior probability is then used to optimise the outcome [Altman et al., 1977]; [Frydman et al., 1985]. However, when the base rate in a sample is very different from the prior probability in the population, this leads to a choice-based sample bias of both the parameter and probability estimate [Manski and Lerman, 1977]. A better solution to this problem is to utilise an adjustment such as the weighted exogenous sample maximum likelihood (WESML) procedure [Zmijewski, 1984]; [Dopuch et al., 1987]; [Manski and Lerman, 1977]. The essence of WESML is modifying the sampling maximum likelihood estimator by weighing each observation's contribution to the log-likelihood. In other words, this procedure incorporates the ratios of proportion of bankruptcy both in the population and in the sample into the maximum likelihood function under an estimation process. Unfortunately, the WESML procedure can not be applied to neural networks since they have no so-called maximum likelihood function. To solve this predicament, this study proposes a solution called weighted cutoff point (WCOP) solution to make the adjustment possible not only in ANNs but also in any other technique. By using this formulation, the choice-based sampling bias will be minimised even when a one to one matching criterion is used in the sample compared with a small percentage bankruptcy rate in real life. Hence, we can, on the one hand, take advantage of eliminating the industry or size effect, and on the other hand, we can avoid the disadvantage of estimation bias.



- (3) To perform a sensitivity analysis of optimal cutoff points to different misclassification costs of Type I and Type II errors.

One criticism of bankruptcy prediction models is that the cutoff point is determined without considering the loss functions of Type I and Type II errors. As a result the classification accuracy is based on the minimisation of the total error probabilities, not the total error costs. It is consequential that the error costs should be taken into account because the misclassification cost of Type I errors is, in general, much higher than that of Type II errors (from the bank's point of view). Although some studies incorporated several different error cost ratios into the model to evaluate the change of error rates, they failed to provide a systematic method to derive an optimal cutoff point which minimises the total error costs. Therefore, the impact of the change of optimal cutoff points to misclassification costs on decision making cannot be fully understood. Through the proposed approach, the optimal cutoff points, accounting for the asymmetrical loss functions of Type I and Type II errors, are mathematically derived. The sensitivity analysis is then conducted in order to observe the robustness of optimal cutoff point and predictive ability to various misclassification costs.

- (4) To test the effect of different compositions of bankruptcy to nonbankruptcy between training and testing data sets.

This involves the problem that if a classification model is built using a training sample with a certain base rate (the ratio of bankruptcy to nonbankruptcy in sample), then the model may not work well when the proportions of bankruptcy to nonbankruptcy in the population has changed, reflecting the fact that the composition of historical data needed in the prediction model cannot be controlled.

## **1.4 The Research Framework and the Key Research Questions**

This study attempts to develop a better understanding of bankruptcy prediction models in terms of horizontal and vertical dimensions. From the horizontal perspective, the thesis examines the capacity of distinct discriminating techniques which involve traditionally used

tools and newly invented ANNs. Its novelty lies in promoting their classification performance on a theoretical basis through a simulation study. Moreover, the introduction of an advanced ANN algorithm in this thesis is different from the common GDR solution of previous studies, and also offers an original approach. From the vertical perspective, apart from validating the simulation results through empirical study, consideration is also given to the possible problems faced by prediction failure, such as the impact of sample size, choice-based design bias, sensitivity analysis of optimal cutoff points to misclassification costs of Type I and Type II errors, and the generalisation ability, all of which have been neglected in previous studies. We not only evaluate the influence of these factors on classification accuracy, but also build a model based on considering the variations of bankruptcy models in reality in order to minimise the bias caused by ignorance of these factors. This will be another significant contribution of this thesis. The following figure displays the framework of the research.

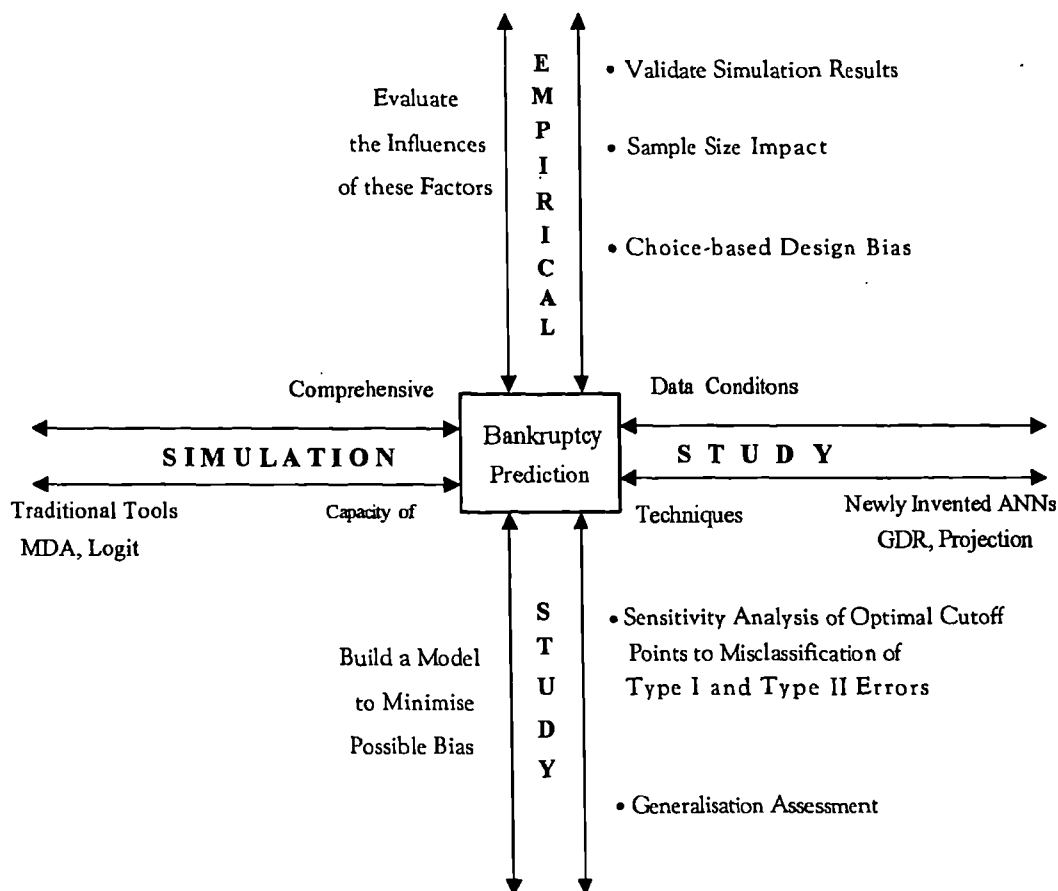


Figure 1.3.1 The Framework of This Thesis

The hypotheses to be tested in this thesis are divided into eight subsets. The first sets compare the classification performance of four alternative techniques. These comparisons are evaluated for various data distributions, group dispersions and orientation schemes individually.

The following hypotheses are proposed

H<sub>1</sub>: There is no difference in predictive ability for normal distribution data among four discriminating methods

H<sub>2</sub>: There is no difference in predictive ability for skewed distribution data among four discriminating methods

H<sub>3</sub>: There is no difference in predictive ability for the data with outliers among four discriminating methods

H<sub>4</sub>: There is no difference in predictive ability for the data with equal variance-covariance matrices across groups among four discriminating methods

H<sub>5</sub>: There is no difference in predictive ability for the data unequal variance-covariance matrices across groups among four discriminating methods

H<sub>6</sub>: There is no difference in predictive ability for the data with high intercorrelation between the two predictor variables among four discriminating methods

H<sub>7</sub>: There is no difference in predictive ability for the data with low intercorrelation between the two predictor variables among four discriminating methods

H<sub>8</sub>: There is no difference in predictive ability for the data with different orientation schemes between the two predictor variables among four discriminating methods

The second hypotheses involve testing the main effects of the three factors (i.e., distribution, group dispersion and orientation) for each of the four methods. The relevant hypotheses are stated as follows

H<sub>9</sub>: The classification performance for each of the four discriminating techniques depends on the data distribution.

H<sub>10</sub>: The classification performance for each of the four discriminating techniques depends on the feature of group dispersion across groups.

H<sub>11</sub>: The classification performance for each of the four discriminating techniques depends on the orientation scheme between two attributes of the data.

The third set to be tested involves the assessment of the possible interaction effects of factors on predictive capability for each of four methods.

The related hypotheses are stated in the following

H<sub>12</sub>: The classification performance for each of the four discriminating techniques is not affected by the interaction effect of data distribution and group dispersion.

H<sub>13</sub>: The classification performance for each of the four discriminating techniques is not affected by the interaction effect of data distribution and orientation schemes of two attributes.

H<sub>14</sub>: The classification performance for each of the four discriminating techniques is not affected by the interaction effect of group dispersion and orientation schemes between two predictor variables.

H<sub>15</sub>: The classification performance for each of the four discriminating techniques is not affected by the interaction effect of data distribution, group dispersion and orientation schemes between two predictor variables.

The fourth set of hypotheses are concerned with the classification accuracy of the four methods using real financial data as well as comparison of simulation conclusions and empirical conclusions.

Four relevant hypotheses are proposed for the experiment carried out in this section.

H<sub>16</sub> : There is no difference in classification performance for four alternative techniques for a training sample based on real financial data.

H<sub>17</sub> : There is no difference in classification performance for four alternative techniques for a testing sample based on real financial data.

H<sub>18</sub> : There is no significant inconsistency between simulation results and empirical results for the four alternative techniques.

H<sub>19</sub> : There is no significant difference in the relative contribution of predictor variables for four alternative techniques

The fifth set of hypotheses test the effects of sample size on predictive ability for the four methods. Whether the statistical methods or neural network are robust to different sample sizes is also an issue here. The following three hypotheses are proposed

$H_{20}$  : The rate of misclassification for each of the four discriminating techniques is not affected by the different sample size.

$H_{21}$  : There is no significant difference in predictive performance among the four alternative techniques for different levels of sample size.

$H_{22}$  : The neural networks are not more robust than the statistical discriminating methods in predictive performance for different sample sizes.

The sixth test sets examine the choice-based sampling issue by using both unadjusted and our proposed adjusted procedure.

The relevant hypotheses in this experiment are stated as follows

$H_{23}$  : The Type I, Type II and Overall error rate has no functional relationship with the decreasing choice-based sample frequency rate in each method when using the unadjusted procedure.

$H_{24}$  : If choice-based sample bias exists, the bias does not decrease when the proportion of the two groups in a sample approaches the prior probability in the population.

$H_{25}$  : The Type I, Type II and Overall error rate has no functional relationship with the decreasing choice-based sample frequency rate in each method when using WCOP procedure.

The seventh test sets are related to the sensitivity analysis of various misclassification costs ratios. The corresponding hypotheses are

$H_{26}$  : The optimal cutoff points that minimise the expected total error costs are insensitive to different misclassification costs of Type I and Type II errors for MDA method.

$H_{27}$  : The optimal cutoff points that minimise the expected total error costs are insensitive to different misclassification costs of Type I and Type II errors for Logit method.

$H_{28}$  : The optimal cutoff points that minimise the expected total error costs are insensitive to different misclassification costs of Type I and Type II errors for GDR method.

$H_{29}$  : The optimal cutoff points that minimise the expected total error costs are insensitive to different misclassification costs of Type I and Type II errors for Proj method.

The eighth question sets concern the effect in terms of differences in base rate between the training sample and the testing sample.

The related hypotheses are indicated below

$H_{30}$  : There is no difference in predictive capability using different base rate between training and testing data composition in MDA technique.

$H_{31}$  : There is no difference in predictive capability using different base rate between training and testing data composition in Logit technique.

$H_{32}$  : There is no difference in predictive capability using different base rate between training and testing data composition in GDR technique.

$H_{33}$  : There is no difference in predictive capability using different base rate between training and testing data composition in Projection technique.

$H_{34}$  : The statistical methods perform as well as the neural networks when the base rate between the training sample and the testing sample is different.

$H_{35}$  : The statistical methods are more robust than the neural networks to different base rates between the training sample and the testing sample is different.

## **1.5 Organisation of the Thesis Chapters**

The thesis consists of eleven chapters. Chapter 1 introduces the objective of this study, discusses emerging trends in classification techniques, and outlines the subjects and methodology to be explored together with the main hypotheses of the research .

Chapter 2 presents two statistical prediction models: multivariate discriminant analysis (MDA) and Logit procedure. Their theories, assumptions, advantages and disadvantages, and the comparisons of these two techniques will be discussed in detail.

Chapter 3 focuses on the artificial neural network technique, including its history, components, characteristics and the learning algorithm. A new learning algorithm, Projection approach, will be introduced to remedy the drawback of the commonly used generalised delta rule (GDR) of backpropagation neural network.

Chapter 4 is concerned with the comparison of statistical and artificial neural network models. In addition to the comparison in theory, the similarities and differences between them reported in previous empirical studies will be also presented.

Chapter 5 discusses the trade-off between multicollinearity and factor analysis. It introduces the research methodology and considers how to generate predictor variables in the simulation study and how to choose the final financial ratios in the empirical study.

Chapter 6 presents the simulation study methodology. Experimental design, data sets generation and statistical analysis instrument will be described in detail.

Chapter 7 analyses the results of four techniques in various levels of data distributions, relationship of variance-covariance matrices, and orientation schemes between predictors. The comparisons of classification performance for the underlying four methods are extensively evaluated.

Chapter 8 discusses the problems encountered with bankruptcy prediction in an empirical estimation process in real life. These include the selection of independent variables, choice-based sample design, cutoff point determination to relative misclassification costs, and generalisation ability. A proposed solution will be developed for each problem, which may then be applied to the empirical study.

Chapter 9 describes the methodology of the empirical study including data collection, variables selection, and research design. The applications associated with each subject to be studied will be carefully handled in order to assess the impact of the factors concerned.

Chapter 10 reports the empirical results corresponding to each highlighted topic. The outcomes will be carefully analysed in order to work out the impacts of experimental factors. The conclusions are expected to be useful for decision makers.

Chapter 11 provides overall comments, the theoretical and empirical contributions implied in this research, the limitations of the study and recommendations for further research.

## Chapter Two

### CONVENTIONAL STATISTICAL METHODS IN BANKRUPTCY PREDICTION

#### 2.1 Linear Discriminant Function

Since Altman [1968] developed the corporate bankruptcy prediction model using multivariate discriminant analysis (MDA) methodology, the MDA has become the most widely used method for identifying financial distress. The discriminant function he employed was Fisher's linear discriminant function (LDF) [Fisher, 1936]. The objective of a multivariate discriminant analysis is to classify observations by a set of independent variables into one of two or more mutually exclusive and exhaustive categories. Here we will focus on the case of two populations. Our problem is to classify an observation (company) into one of two categories (bankruptcy and nonbankruptcy) based on a vector of characteristics  $\mathbf{X} = (x_1, x_2, \dots, x_n)$ .

Let each observation's discriminant score  $Z_i$  be a linear function of the independent variables  $\mathbf{X}_i$ . It means,

$$Z_i = b_0 + b_1X_{i1} + b_2X_{i2} + \dots + b_nX_{in} = \mathbf{b}'\mathbf{X}_i$$

where

$Z_i$  = the  $i$ th observation's discriminant score

$X_{in}$  = the  $i$ th observation's value of the  $n$ th independent variable

$b_n$  = the discriminant coefficient for the  $n$ th independent variable

The discriminant function separates the observation in a linear way. The classification boundary is the locus of points  $b_0 + b_1X_{i1} + b_2X_{i2} + \dots + b_nX_{in} = Z^*$  ( $Z^*$  is the optimal cutoff point for the discriminant score). When  $n=2$ , the classification boundary is a straight line. When  $n=3$ , the classification boundary is a two-dimensional plane in three-dimensional space, and thus the  $n-1$  dimensional hyperplane in a  $n$ -dimension space. The LDF maps points, representing observations in two different categories, from a  $n$ -dimension attribute space into an one-dimensional space in such a way that the distributions of points from the



groups are maximally separated. The algorithm to generate the classification criterion involves using linear combinations of predictor variables and choosing the coefficients so that the ratio of the squared difference between  $\bar{Z}_1$  and  $\bar{Z}_2$  (the means of  $Z$  in the two group) to the variance of  $Z$  is maximised. That is, Fisher's approach advocated the maximisation of the ratio of the among-groups sum of squares on the function to the pooled within-groups sum of squares on the function [Cooley and Lohnes, 1971]. Let the means of  $X$  vector in the two groups be  $\mu_1$  and  $\mu_2$ , and the covariance matrices of  $X$  in the two groups be  $\Sigma_1$  and  $\Sigma_2$ , respectively. Thus the means of the linear function  $Z$  in the two groups are  $b\mu_1$  and  $b\mu_2$  ( $b$  is the vector of discriminant coefficients  $b_i$ s). If we assume  $\Sigma_1 = \Sigma_2 = \Sigma$ , then the variance of  $Z$  is  $b\Sigma b$ . Thus, we need to maximise

$$\phi = \frac{[b(\mu_1 - \mu_2)]^2}{b\Sigma b} = \frac{\text{Between Group Variance}}{\text{Within Group Variance}} \quad (2.1.1)$$

Differentiating (2.1.1) with respect to  $b$  and equating the derivative to zero, we get

$$b = \Sigma^{-1}(\mu_1 - \mu_2)$$

Generally, the parameters  $\mu_1$ ,  $\mu_2$  and  $\Sigma^{-1}$  are not known. It is usual practice to estimate them by the corresponding sample mean  $\bar{X}_1$  and  $\bar{X}_2$ , and sample variance  $S^{-1}$ . Thus, the means of the discriminant functions in the groups can be expressed respectively

$$\begin{aligned} \bar{Z}_1 &= b' \bar{X}_1 = (\bar{X}_1 - \bar{X}_2)' S^{-1} \bar{X}_1 \\ \bar{Z}_2 &= b' \bar{X}_2 = (\bar{X}_1 - \bar{X}_2)' S^{-1} \bar{X}_2 \end{aligned}$$

Given a new observation with characteristics  $X_0$ , then the output  $Z_0$

$$Z_0 = b' X_0 = (\bar{X}_1 - \bar{X}_2)' S^{-1} X_0$$

$X_0$  is assigned to the first group, if  $Z_0$  is closer to  $Z_1$  than  $Z_2$ . Assuming  $\bar{Z}_1$  is greater than  $\bar{Z}_2$ ,  $Z_0$  will be closer to  $\bar{Z}_1$  than to  $\bar{Z}_2$  if

$$|Z_0 - \bar{Z}_1| < |Z_0 - \bar{Z}_2|$$

It means

$$Z_0 > 1/2 (\bar{Z}_1 + \bar{Z}_2)$$

Thus, the optimal cutoff point  $Z^*$  is the average of the two means. The square of the difference between the means is often called the Mahalanobis generalised distance and is denoted by  $D^2$

$$D^2 = (\bar{Z}_1 - \bar{Z}_2)^2 = (\bar{X}_1 - \bar{X}_2)' S^{-1} (\bar{X}_1 - \bar{X}_2)$$

Up to now, we are not making the assumption of normality. However, to apply any test of significance, the assumption of normality is needed. The explanatory variables in the two groups should come from normal populations with means  $\mu_1$  and  $\mu_2$  respectively, and the same covariance matrix  $\Sigma$ . Under this assumption, the F ratio is used to test whether or not there are significant differences between the two groups

$$F = \frac{n_1 n_2 (n_1 + n_2 - k - 1)}{(n_1 + n_2) (n_1 + n_2 - 2) k} D^2$$

where

F ratio with degrees of freedom  $k$  and  $(n_1 + n_2 - k - 1)$   
 $k$  is the number of explanatory variables.

This is known as Hotelling's  $T^2$  test for the hypothesis  $\mu_1 = \mu_2$ , assuming normality for the distribution of  $X$  and a common covariance matrices  $\Sigma$ . The method outlined here in deriving  $b$  is distribution free, but the F test is not.

In essence, Fisher's linear function requires that the predictor variables are normally distributed, and that the populations have equal variance-covariance matrices. Otherwise, a linear classification rule and the test for distinction between groups, i.e., the Hotelling's  $T^2$ , become inappropriate.

Further, the optimal cutoff point  $Z^* = 1/2(\bar{Z}_1 + \bar{Z}_2)$  can be improved on if we have some prior probabilities that  $X$  belongs to either population, and if we are given the misclassification costs of the two type errors.

## 2.2 Problems with Discriminant Analysis

Fisher developed the above linear discriminant function and dealt with the problem of correctly classifying iris plants into one of two populations, iris setosa and iris versicolor, based on the length and width measurements of the sepal and petal. This approach derived a composite score for each observation by choosing the coefficient  $\mathbf{b}$  so that the variance of  $\mathbf{b}'\mathbf{X}$  between groups is maximum relative to its variance within groups. Although in his original derivation, Fisher used a linear regression approach that did not require any distributional assumptions for independent variables, the classification procedure ability is effective only with the multivariate normal distribution in independent variables, and the equal variance-covariance matrices across the groups. Specifically, the linear discrimination function can be justified either by a least squares argument or by assuming multivariate normality [Lee and Ord, 1990]. Simultaneously the spreads of the independent variables (the  $\mathbf{X}$ 's) in group 1 are assumed the same as the spreads in the independent variables in group 2, and the interrelations (correlation) among the independent variables in groups 1 are also assumed to be the same as the interrelations in group 2. When these conditions are satisfied, and the prior probabilities in population for both groups are known, then the MDA provides the optimal classification rule and the discriminant coefficients are the true maximum-likelihood estimates (MLE) of the discriminant function. Otherwise these estimates are neither efficient nor consistent. Unfortunately, these restrictive assumptions are often violated when real world financial data is used. In practical bankruptcy prediction there is no possibility that the financial ratios which explain the reasons for financial distress are multivariate normally distributed and equally dispersed across groups. Furthermore, when the prior probabilities of two groups in the population and the two kinds of misclassification costs are considered, and if the objective is to minimise the expected total cost of misclassification, a cutoff point  $Z^*$  will be changed to

$$Z^* = \log \frac{(1 - \alpha_p) C_I}{\alpha_p C_{II}}$$

where

$\alpha_p$  = the prior probability of a bankrupt group

$C_I$  = the cost of misclassifying an observation as belonging to a nonbankrupt group when it actually belongs to a bankrupt group

$C_{II}$  = the cost of misclassifying an observation as belonging to a bankrupt group when it actually belongs to a nonbankrupt group

( A theoretical argument was presented in Anderson [1958] and Morrison [1969] )

This implies that the appropriate cutoff point value depends on both the prior probability of group membership and the ratio of costs of misclassification. In other words, cutoff points used in the previous reports of financial applications would not be optimal without assuming prior probabilities identical to sample group frequencies and the equality of  $C_I$  and  $C_{II}$ . However, these conditions are rarely satisfied. Accordingly, the conclusions and generalisations that can be drawn from some previous studies may be tenuous and questionable. More importantly, as we can see from the implication of the derivation of discrimination function, MDA requires that the decision set used to distinguish between a failing group and a nonfailing group must be separable in linear terms. This implication, as well as the stringent assumptions, makes the MDA incompatible with the complex nature, boundaries and the interrelationships of financial ratio in failure prediction. Consequently, the power of MDA for financial distress analysis is compromised and the results may be unreliable [Karels and Prakash, 1987].

## **2.3 Studies Evaluating the Performance of MDA**

### **2.3.1 Normality Assumption**

Horrigan [1965]; Mecimore [1968]; O'Connor [1973]; Deakin [1976]; Bougen and Drury [1980]; and Karels and Prakash [1987], etc. have tested the basic hypothesis of normality in financial data. Horrigan [1965] analysed seventeen ratios for 50 USA companies over the period 1948-1957 and suggested that most financial ratios tended to be normally distributed but that there was some evidence of positive skewness. O'Connor [1973] also analysed ten ratios for 127 companies in the USA but for the different period covering 1950-1960. He found that although most ratios distributions were skewed, the central area of the distribution was approximately symmetrical. Subsequently, the cross-sectional distribution of eleven ratios over the period 1955-1973 covering 1800 companies of USA manufacturing firms was comprehensively investigated by Deakin [1976]. He concluded

that the normality assumption was untenable for these eleven well-known ratios, except for the debt/total asset ratio. Similarly, Bougen and Dury [1980] explored the distributional properties of seven financial ratios for over 700 UK companies in 1975. The overall results suggested non-normality both for the whole sample and at individual industry level.

Questions have been raised in the literature about the success of MDA in predicting firm bankruptcy due to the violation of multivariate normality. There were several examinations of the robustness of the MDA to the non-normality. Gilbert [1968] reported that discriminant analysis may be robust to normality violations. Lachenbruch, Sneeringer and Revo [1973] indicated that linear model including MDA was reasonably robust but can be sensitive to heavy tails or outliers in the data, and suggested the data should first be transformed to approximate normality. In the presence of outliers, in particular, they found that the results of discriminant analysis may be seriously misrepresented. The error rate for certain groups (i.e., Type I error or Type II error) can be distorted even if the Overall error rates are not significantly affected. Thunhurst [1985] also confirmed that the models were sensitive to the presence of skewed data and extreme values, and that the discriminant function can be dominated by few very large observations which may significantly reduce its usefulness for decision-making purposes. Chinganda and Subrahmaniam [1979] investigated the robustness of the MDA and concluded that where possible one should first attempt to transform the feature data to normality before constructing the MDA model. The usual solution for coping with this problem is to apply the natural or standard log or square roots transformations [Carleton and Lerner, 1969]; [Horton, 1970]; [Pinches and Mingo, 1973]; Bates [1973], which was suggested by Kirk [1968] in order to make data fit more closely to the normal distribution. These common procedures has been applied not only to solve the violation of non-normality in quantitative continuous independent variables, but also especially in qualitative variables such as firm sizes, industry index and macroeconomic variables.

In spite of the fact that log transformations, square root transformations, and winsorizing (changing an outlier's value to that of the closest non-outlier, and then attempting to fit the distribution with a known one) and trimming (segregating outliers) approaches have often been cited to mitigate the effect of non-normality, in reality financial ratios have generally not been successively transformed in the bankruptcy prediction literature. This is not surprising, since financial ratios are constructed from two accounting variables, and the

joint distribution will depend on the behaviour of both the numerator and the denominator and on the relationship between those two co-ordinates. If there is non-proportionality, the distribution will be skewed [Barnes, 1982]. This non-proportionality probably explains why even after transforming data or eliminating outliers, normality could still not be achieved [Ezzamel, Mar-Molinero and Beecher, 1987]; [Lee, 1985]. Additionally, both natural logs and square roots suffer from the defect that they cannot be applied if the ratios are negative. Although other alternatives which avoid this difficulty are available (for example adding or subtracting a constant or taking squares of the ratio), such methods tend to accentuate the distortion caused by outliers giving more weight to large observation [Ezzamel et al., 1987]. Watson [1990] also cautioned that the use of transformation techniques may not preserve the statistical properties of the financial ratios and may even change the interrelationships among the variables. Eisenbeis [1977] stated this concern using an example that if the variable being log transformed was that of firm size, the implication would be that the difference between a \$1 billion firm and a \$2 billion firm is less than that between a \$1 million firm and a \$2 million firm, since the percentage difference in the log will be greater in the latter case than in the former. The facts that the transformed variables give less weight to equal percentage changes in a variable when the values are larger than when they are smaller may lead to confusing results.

Zhezhe [1968] undertook a comparative performance study to assess the impact of non-normality on classification accuracy. Various linear discriminant functions containing Fisher's LDF and using both arbitrary distributions as well as normal distribution were investigated while assuming equal variance-covariance matrices across groups. The results indicated that the classification accuracy would be better if applying to multivariate normal variates than to non-multivariate normal variates. Chinganda and Subrahmaniam [1979] also reported on an experiment about the effect of non-normality on errors of misclassification, and found that "when the overlap is excessive, the experimental is best off making a normalising transformation. Otherwise there is a considerable probability of misclassification" (p.76). Karels and Prakash [1987] directly tested the assumption of multivariate normality on the financial ratios used in previous studies. Results indicated that there is no evidence these ratios were from a multivariate normal population. They then chose some of the least non-multivariate normal financial ratios in order to test predictive ability. It was shown that the discriminant model obtained using these

discriminators considerably improved the forecasting accuracy. Gessner et al. [1988] examined the effects of three assumption violations (1) non-normal joint distribution of the independent variables; (2) unequal group variance-covariance matrices; (3) multicollinearity of the independent variables for five estimation techniques including linear discriminant analysis. The empirical results in this article clearly showed that if the assumptions underlying dichotomous dependent variable prediction are violated, parameter estimates may be misleading, even when the goodness of fit statistics is not substantially affected. Thus, the discriminant results provided by Altman [1968, 1974]; Deakin [1972]; Blum [1974]; Edmister [1972]; Levitan and Knoblett [1985]; and etc. may be suspect.

### 2.3.2 Equal Group Dispersion Assumption

Another important assumption in the application of LDF is that the variance-covariance matrices are equal across all groups. Eisenbeis [1977] suggested that if this assumption is not fulfilled, it will not only affect the significance test for the equality of group means but also the appropriate form of the classification rules. The importance of this attribute can not be easily described using the tediously mathematical computation. But a simple example may reveal the possible impact of the violation of this assumption on classification accuracy in using MDA.

Suppose  $\sigma_{ij}^1$  denotes the covariance between variables  $i$  and  $j$  for observations belonging to group 1 and the covariance matrix denoted  $\Sigma_1$  is

$$\Sigma_1 = \begin{bmatrix} \sigma_{11}^1 & \sigma_{12}^1 & \bullet & \bullet & \bullet & \sigma_{1n}^1 \\ \sigma_{12}^1 & \sigma_{22}^1 & \bullet & \bullet & \bullet & \sigma_{2n}^1 \\ \bullet & \bullet & & & & \bullet \\ \bullet & \bullet & & & & \bullet \\ \bullet & \bullet & & & & \bullet \\ \sigma_{n1}^1 & \sigma_{n2}^1 & \bullet & \bullet & \bullet & \sigma_{nn}^1 \end{bmatrix}$$

The mean vector  $m$  in group 1 is denoted as

$$\mu_1 = (\mu_{11}, \mu_{21}, \dots, \mu_{n1})$$

where  $\mu_{ij}$  = the  $i$ th variable for group 1

Analogous definitions hold for the mean vector  $\mu_2$  and covariance matrix  $\Sigma_2$ .

The covariance between  $X_i$  and  $X_j$  is equal to the covariance between  $X_j$  and  $X_i$ . That is  $\sigma_{ij}^1$  is equal to  $\sigma_{ji}^1$  and the matrix is symmetrical.  $\sigma_{ii}^1$  is just the variance of  $X_i$ . The correlation coefficient represents the simple linear correlation between two variables. It can be expressed

$$r_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}} \sqrt{\sigma_{jj}}}$$

For simplicity, a case of two groups on the basis two variables  $X_1$  and  $X_2$  is to be classified. If the covariance of group 1 and group 2 are of the forms respectively

$$\Sigma_1 = \begin{bmatrix} \sigma_{11}^1 & 0 \\ 0 & \sigma_{22}^1 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} \alpha\sigma_{11}^1 & 0 \\ 0 & \alpha\sigma_{22}^1 \end{bmatrix}$$

where  $\alpha > 1$ . However, the mean vectors  $\mu_1$  and  $\mu_2$  are assumed equal. The common mean vector is denoted as  $\mu$ . Obviously, the farther an individual's  $X$  is from the common mean vector  $\mu$ , the more likely it is that it is from group 2.

Mathematically, we would calculate the distance from  $\mu$  at which the likelihood functions for each group were equal. Because the covariance matrices are symmetrical, the locus of such points will be a circle with  $\mu$  as the centre. The classification boundary will be this circle. That is, these unequal variance-covariance matrices can lead to a non-linear classification boundary, and the classification accuracy can not be correct using the linear discriminant function.

The other impacts of violation of this attribute were presented by Cooley and Lohnes [1962, 1971]; Rulon et al. [1967]; and Tatsuoka [1971]. As we know, one of the advantages of discriminant analysis is its reduction in dimensionality, which can be used to reduce the original  $n$  dimensional variable test space to a one-dimensional problem. However, it has been proved that this reduction in dimensionality can leave the significance tests and classification results unaffected if and only if the group dispersion matrices are equal. If group dispersions are not equal, then the transformation to reduced space is no longer distance perserving. It will confound the relative positions of the observations in reduced space, and thus affect the significance test.



This problem arises because the variation of the financial ratios of failed firms is likely to be very different than that of successful firms. Heterogeneity of variance-covariance matrices is a virtually inescapable fact, hence the dispersions of two different groups can not be pooled or combined as was the same procedure with Fisher's LDF.

Many studies have indicated that Fisher's LDF was not robust to the violation of the assumption of homogeneity of the variance-covariance matrices. Smith [1947] developed this procedure using multivariate normal but unequal variance-covariance data within groups. When the two groups bivariate data were graphed, the boundary between the two groups was a quadratic figure rather a line. This is why Smith's discriminant function is well known as quadratic discriminant function (QDF).

Gibert [1969] examined the effect of unequal variance-covariance matrices. She learned that there was almost no difference in the predictive ability of LDF and QDF when the data has just a slight inequality of group dispersion under multivariate normal population, and when group overlap is small. However, as the number of variates increased, the QDF became superior.

Marks and Dunn [1974] have generated two multivariate normal populations with unequal variance-covariance matrices, assuming equal misclassification costs, but varying the number of variates, sample size and a prior probabilities. They reached similar conclusions, that if the variance-covariance matrices are not equal and the sample size is sufficiently large relative to the number of predictive variables, then QDF rather than LDF yielded the optimal solution. However, for a small sample size, especially as the number of predictive (independent) variables decreases, the LDF outperformed the QDF. Wahl and Kronmal [1977] also suggested that QDF performs worse than LDF for small sample sizes. A more comprehensive study was undertaken by Lachenbruch, Sneeringer and Revo [1973]. They evaluated the robustness of both LDF and QDF to the violation of multivariate normality and homogeneity of variance-covariance at the same time. Additionally, the parameters varied in this study were the sample size and the number of variates, while the misclassification costs and prior probabilities were assumed equal. The consequent studies achieved the following results: (1) LDF and QDF are both affected by non-normality, but the QDF is more affected than LDF. (2) Error rates increase as heterogeneity of group dispersion increases; whereas it decreases as the number of variates increases. Joy and

Tolleffson [1975] demonstrated that if the assumption of homogeneous covariance matrices is not met, the significance test for the equality of group means will be affected. In this situation the quadratic discriminant rule instead of linear discriminant rule should be used to minimise the probability of misclassification. Pinches [1980] also indicated that misclassification rates are influenced for both linear and quadratic functions, and that the latter is affected even more than the former. Further, the research by Altman, Haldeman and Narayanan [1977] reported that the QDF may be superior to LDF in the corporate financial distress prediction model, since the financial ratios of bankrupt firms and nonbankrupt firms are not likely to have the equal variances of independent variables and the interrelationships between variables across these two groups.

However, the investigation made by Hamer [1983] provided a different conclusion. She examined four different data sets which were previously employed by Altman [1968]; Deakin [1972]; Blum [1974]; Ohlson [1980] to survey the impact of the assumption of equal group dispersion. She found that for each of the data sets, the linear model performed at least as well as the quadratic version for classification accuracy despite the existence of statistical significant differences in the variance-covariance matrices for failed and nonfailed groups. In general the review of these studies has suggested that if the data is close to normal distributions and has extremely unequal variance-covariance matrices, then the QDF yields a better classification performance than Fisher's LDF, especially for large sample sizes. When the variance-covariance matrices are only slightly unequal, the QDF performs only slightly better. For small sample size and slight heterogeneity, Fisher's LDF is superior to QDF.

## **2.4 Nonparametric Discriminant Functions**

Despite the fact that applying the QDF approach can overcome the assumption of homogeneity of variance-covariance matrices across groups, the attribute measures that discriminate amongst the population should still be jointly multivariate normal. Otherwise, neither linear nor quadratic discriminant analysis procedure will be optimal. Indeed, within almost all research literature on empirical failure prediction, investigation of the statistical

test about both univariate and covariance matrices has demonstrated that both normality and equal group dispersion assumptions are violated. In this case a nonparametric version of discriminant analysis may be an alternative to be utilised. Nonparametric discriminant methods are based on nonparametric estimates of group-specific probability density. Either a kernel method or the k-nearest-neighbor method can be used to generate a nonparametric density estimate in each group and to produce a classification rule. The kernel method requires the assumption of a particular distribution as uniform, normal, biweight, or triweight in the density estimation. While the idea behind the k-nearest-neighbor method is to calculate the smallest differences between a specific observation and other observations within the pooled group.

Tam and Kiang [1990] utilised the k-nearest-neighbor method in their study of the bank failure prediction model. Due to the non-normality of the financial ratios, this nonparametric approach was thus chosen to achieve the classification task, but the performance of this study is not as satisfactory compared to artificial intelligence technique. Dwyer [1992] undertook empirical comparisons of the effectiveness of nonparametric discriminant technique, logistic regression and artificial neural network models in corporate bankruptcy prediction. The results obtained through the use of the nonparametric technique based on k-nearest-neighbor approach were generally disappointing. Overall nonparametric discriminant procedure, which was developed for data attribute's violation of the assumptions of LDF, does not provide good classification capability. Accordingly, it has rarely been adopted by researchers.

## **2.5 The Reason for Selecting LDF in Our Study**

In this thesis linear discriminant analysis is selected as a representative of discriminant functions to make comparative analyses to the other statistical method, -Logit procedure, and two artificial neural networks. The reasons are

1. The LDF has more intuitive appeal and allows a clear interpretation on each of the explanatory variables which can not be isolated by QDF.
2. The QDF is very sensitive to non-normal data in spite of taking a few advantages of heterogeneity of group covariance matrices. There will be more distortion in applying

the QDF rule than the LDF rule if the financial ratios can not be multivariate normally distributed.

3. The QDF usually requires a larger sample size in the estimation process in order to avoid the overfitting problem, which may not be appropriate when the size of sample is one of our concerns in this thesis.
4. The nonparametric discriminant methods have had little popularity in the previous literature. Thus, its results may not be reliable and are difficult to compare with those of other techniques.

## 2.6 The Importance of Prior Probabilities and Misclassification Costs

In constructing a bankruptcy prediction model, optimal classification criteria have often been determined by minimising the total number of misclassifications under the assumption of equal prior probabilities and misclassification costs for two groups. The results from this model will lead to a large number of Type I errors when used in real life because nonbankruptcy occurs much more frequently than bankruptcy. On the other hand, merely considering prior probabilities, and ignoring the inequality of misclassification costs of Type I and Type II errors, the results are still not optimal because the cost of Type I error can be much higher than that of Type II error.

Joy and Tollefson [1975] has revealed the bias when prior probability is not taken into account in a decision making context. The well-known study developed by Altman [1968] is presented here in order to demonstrate this problem. The cross-validation results from Altman are shown below

Table 2.6.1 Altman Cross-Validation Results

Actual group membership	Predicted Group Membership		
	Bankrupt	Nonbankrupt	Total
Bankrupt	24 ( $n_{11}$ )	1 ( $n_{12}$ )	25 ( $n_{1.}$ )
Nonbankrupt	14 ( $n_{21}$ )	52 ( $n_{22}$ )	66 ( $n_{2.}$ )
Total	38 ( $n_{.1}$ )	53 ( $n_{.2}$ )	91 ( $n_{..}$ )

In Altman's study, the inferential analysis was based on that the assumption the proportions of two groups in the sample is equal to those in the population. That is, the prior probability of bankruptcy  $\alpha_p=25/91=0.275$ , and the prior probability of nonbankruptcy is  $(1-\alpha_p)=66/91=0.725$ . On this basis the total classification accuracy is  $(24+52)/91=(0.275)(24/25)+(0.725)(52/66)=0.835$ . On the other hand, according to the proportional chance model, under which entities are randomly assigned to groups with probabilities equal to group frequencies, and which implies that prediction by guessing can achieve a correct rate for each group involved equal to the proportion of that group [Huberty, 1984], the expected fraction of correct classifications under this scheme for the Altman study is then  $(25/91)(25/91)+(66/91)(66/91)=0.61$ . Put another way, the Altman's model is better than a chance classification standard under the above assumptions.

However, if the prior probability is assumed to be 0.02, as utilised in another Altman study [1977], the estimated total expected fraction of correct classification for Altman's LDF becomes  $(0.02)(24/25)+(0.98)(52/66)=0.791$ . This outcome will be worse than the 0.96  $((0.02)^2+(0.98)^2)$  classification accuracy obtained from the proportional chance model. This is the result of the prior probability not being considered in establishing the cutoff point. Thus,  $Z^*$  that is optimal for classification of a sample with proportions  $\alpha_p=0.275$  (bankrupt) and  $1-\alpha_p=0.725$  (nonbankrupt) will not be optimal for classification of a sample with the  $\alpha_p=0.02$  and  $1-\alpha_p=0.98$ . Furthermore, in terms of another interesting conditional efficiency—the Bayesian posterior conditional probability, which measures the probability a firm is actually a bankrupt (nonbankrupt) given a bankrupt (nonbankrupt) classification, the probability of bankruptcy given by Altman's classification of bankrupt will be calculated as

$$\frac{(n_{11}/n_1)q_1}{(n_{11}/n_1)q_1 + (n_{21}/n_2)q_2} = \frac{(24/25)q_1}{(24/25)q_1 + (14/66)q_2}$$

The result obtained is only 0.085 when the prior probabilities of bankruptcy is 0.02( $q_1$ ), down from 0.632 when the prior probabilities and sample proportions were assumed to be identical. This dramatic difference underscores the importance of properly incorporating prior probabilities in the analysis.

Jones [1987] stated that when the prior probability of bankruptcy is much lower than that of nonbankruptcy, the optimal cutoff point will be adjusted by moving away from the

mid-point between group means and closer to the failed firms mean. The tendency is to favour nonfailed group resulting in misclassifying more failed firms into the nonfailed group, while misclassifying less nonfailed firms into the failed group, and the overall classification is then improved because of the large proportion of nonfailed firms in the sample. Jones concluded [1987] that failure to consider unequal prior probabilities of two populations is a valid criticism of earlier studies.

In addition to prior probability, the cost of misclassification costs should also be assessed in evaluating predictive ability. Misclassification costs are mostly subject to the subjective judgement associated with specific consideration, and may not be easily estimated. However, one would expect that the Type I error cost is greater than Type II error cost [Altman, 1980], [Hsieh, 1993] (see section 8.4 for details). The impact of misclassification costs of Type I and Type II errors on determining optimal cutoff points has not attracted much attention in previous research. This ignorance may result in a non-optimal bankruptcy prediction model since the higher the misclassification cost of Type I error to that of a Type II error, the more the Type I errors should be minimised, especially when the cutoff points are very sensitive to these misclassification costs. Therefore, misclassification costs are an important factor and should be considered in building the model.

## **2.7 Logit Method and Conditional Probability Approach**

### **2.7.1 Logistic Regression**

To avoid the assumptions of MDA, Ohlson [1980] made a change in the discriminating techniques for evaluating business failure prediction models in his study. He used a logistic regression (Logit) procedure instead of the traditional MDA methodology. The Logit procedure is also a statistical method, which however places neither restrictions on the distribution of independent variables, nor on the structures of the covariance matrices of two groups. Additionally, a Logit analysis provides the probability that an observation will fall within a given group while an MDA composite score has little interpretative meaning as it is simply an order-ranking device. The Logit is actually one of a class of linear

conditional probability models. The model assumes that there is an underlying latent dependent variable,  $Z$  defined by the following regression relationship

$$Z^*_i = \sum_{j=0}^n b_j X_{ij} + e_i = \mathbf{b}'\mathbf{X}_i + e_i \quad (2.7.1)$$

where

$Z^*_i$  = the response variable defined by the regression relationship for observation  $i$

$b_j$  = coefficients of independent variables  $X_j$ , the unknown model parameters

$X_{ij}$  = the  $i$ th observation's value of the  $j$ th independent variables

$e_i$  = error term for observation  $i$ . It is an independent and identically distributed random variable with mean zero.

In practice,  $Z^*$  is unobservable. For a binary classification problem, this variable is related to an observable dummy variable  $Y$  through the relation

$$\begin{aligned} Y &= 1, \text{ if the firm is financially distressed or } Z^*_i > 0 \\ Y &= 0, \text{ otherwise} \end{aligned} \quad (2.7.2)$$

From the equation (2.7.1) and (2.7.2), it is clear that:

$$\text{Prob}(Y_i=1) = \text{Prob}(\varepsilon_i > -\beta'\mathbf{X}_i) = 1-F(-\beta'\mathbf{X}_i)$$

where  $F$  = the cumulative distribution function for  $e$ .

In effect the observed values of  $Y$  are simply realisations of a binomial process with probabilities defined by equation (2.7.1) and varying from trial to trial depending on  $\mathbf{X}_i$ . Hence the likelihood function  $L$  of observing  $Y_i$ s can be expressed as

$$L = \prod_{Y=0} F(-\beta'\mathbf{X}_i) \prod_{Y=1} [1-F(-\beta'\mathbf{X}_i)] \quad (2.7.3)$$

The functional form of  $F$  depends on the cumulative distribution of  $\varepsilon$  in equation (2.7.1). This function describes the relationship between the dependent and independent variables and should be continuous and differentiable. Due to the lack of a full theory of bankruptcy,

the best deterministic class of function  $F$  can not be easily found. For the sake of computational and interpretative simplicity, the logistic distribution function is chosen

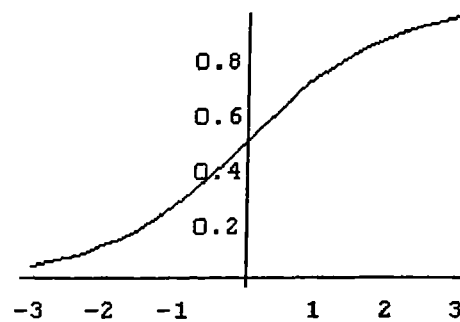
$$P = F(Z) = \frac{1}{1 + e^{-Z}} \quad 0 < P_i < 1 \text{ and } Z_i = \beta'X_i \quad (2.7.4)$$

$P_i$  represents the conditional probability for any given  $X_i$  and  $\beta$ . If the cumulative distribution of  $\varepsilon$  is the logistic, we have the Logit model. In this case

$$F(-b'X_i) = \frac{1}{1 + \exp(\beta'X_i)}$$

$$1 - F(-b'X_i) = 1 - \frac{1}{1 + \exp(\beta'X_i)} = \frac{\exp(\beta'X_i)}{1 + \exp(\beta'X_i)}$$

We observe that there is a closed-form expression for  $F$ , because it does not involve integrals explicitly. Further, the above formula (2.7.4) has two implications. First,  $P$  is increasing in  $Z$ . Second,  $Z$  is equal to  $\log [P/(1 - P)]$ . The model is thus easy to compute and interpret, and is its main virtue [McFadden, 1973]. The Logit model in the present model assumes that  $P_i$  represents the probability of bankruptcy for the  $i$ th company having characteristics  $X_i$ . This predicted probability is mapped to the meaningful zero-one range. The logistic cumulative distribution function (c.d.f.) is a sigmoid curve that asymptotically approaches zero and one. The following figure displays this situation.



**Figure 2.7.1 The Cumulative Distribution of Conditional Probability in Logit Procedure**

In order to estimate the coefficients in the Logit analysis, the maximum likelihood method (MLE) is used. The likelihood function for use in sample estimation of the coefficients of  $Z$



in (2.7.3) is given by multiplying the products of all  $P_i$ s for bankrupt firms times the product of all amounts  $1-P_i$  for all nonbankrupt firms. So higher failure probabilities for failed firms and lower failure probabilities for nonfailed firms represent higher points on the likelihood function. The likelihood function  $L$  can be rewritten as

$$\begin{aligned}
L &= \prod_{i=1}^n \text{Prob}(Y_i = 0) \prod_{i=1}^n \text{Prob}(Y_i = 1) \\
&= \prod_{i=1}^n \text{Prob}(e_i < -b'X_i) \prod_{i=1}^n \text{Prob}(e_i > -b'X_i) \\
&= \prod_{y=0} F(-b'X_i) \prod_{y=1} [1-F(-b'X_i)] \\
&= \prod_{i=1}^n \left( \frac{1}{1+\exp(\beta'X_i)} \right)^{1-y_i} \left( \frac{\exp(\beta'X_i)}{1+\exp(\beta'X_i)} \right)^{y_i} \\
&= \frac{\exp(\beta') \sum_{i=1}^n X_i Z_i}{\prod_{i=1}^n [1+\exp(\beta'X_i)]}
\end{aligned}$$

The coefficients estimate  $b$  of  $\beta$  can be obtained by finding the global maximum of the logarithm of the likelihood function. That is, differentiate the equation and set it equal to zero. Due to the nonlinearity of the partial derivatives in  $\beta$ , an iterative technique such as the Newton-Raphson method must be used to determine this global maximum. The maximum likelihood estimates are considered consistent and asymptotically efficient for large values of  $N$ . Then the  $P$  can be achieved through the estimated parameters. The probability estimate of entering group  $k$  will always be between 0 and 1, regardless of the value of  $Z$ . The observations are thus classified into the group which they have the highest predicted probability of entering, or by comparing a certain predefined threshold level to determine which group then should fall into. Moreover, the probability is that the observation will declare the occurrence of group  $k$ .

Some attributes are worth paying attention to in Logit procedure. Because of the curvilinear nature of the cumulative probability of the logistic function (as seen in Figure 2.7.1), the slope of the curve is steeper in the midrange. It leads to the fact that the midrange of probabilities is more sensitive to changes of value in predictor variables. It means that once the probability of bankruptcy is close to 1, changes in predictor variables are likely to raise the probability only by small amount. Similarly, probabilities close to 0 will not easily be reduced even with significant changes in predictor variables; while a small

unit change in predictor variables is more influential at the midrange of probability [Jones, 1987]. Actually, in predicting bankruptcy problems, the marginal companies, rather than quite healthy and quite distressed companies, may be the key issue for a model's validation. This implies that for the marginal companies which reside in the midrange of probabilities, small variations in independent variable values will easily sway the most important probabilities we can forecast [Jones, 1987].

Collins and Green [1982] also indicated that the logistic functional form has the "threshold" property that the bankruptcy forecasting problem logically requires since the curve asymptotically approaches zero and one. For example, suppose  $X$  is the ratio of debt to asset, then various  $Z$  values computed from  $b'X$  on companies are mapped to the function presented in Figure 2.7.2. Under this circumstance, the logistic cumulative distribution function (c.d.f.) would suggest that the "breaking point" is around 30% as illustrated in Figure 2.6.2. That is, if the quick ratios increases from zero to 30%, no substantial increase in the probability of failure would be predicted, but an increase from 30% to 70% would make the probability increase to almost one. Additionally, a firm with a debt to asset ratio of more than 70% would not have a much higher probability of failure than one with 70%. From the practical viewpoint it is entirely reasonable that some key financial ratios have a powerful distinguishable ability within a certain range, but retain the same property within an other range. Furthermore, if more than one independent variable is used, the model estimates the linear combination of ratios that is more likely to produce a good "breaking point". This is the reason why Collins and Green [1982] posited that Logit has much more theoretical appeal and much better statistical threshold properties as a bankruptcy prediction model.

However, the curvilinear nature of the model makes the interpretation of coefficient values rather complex. A change in the probability of bankruptcy can not be directly estimated by multiplying a coefficient times the changes in the value of an independent variable at all ranges, since the midrange of probability is more sensitive to changes in the values of independent variables. Pindyck and Rubinfeld [1981] demonstrated this situation with the example of a voting study. When the probability of a vote is 20 percent, change in registration (moving the independent variable from a value of zero to one) will increase the probability by 3.8 percent, other things being equal. In contrast, if the probability is 50 percent, changing registration will produce a growth in the probability of 5.5 percent.

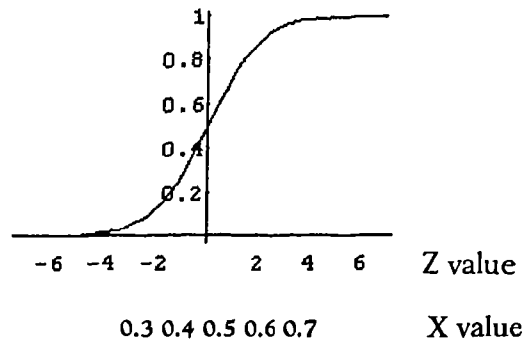


Figure 2.7.2 The Hypothetical Logit Cumulative Density Function

### 2.7.2 Other Link Functions and Corresponding Distributions

As we have stated before, in order to transform the  $Z$  value into the  $[0, 1]$  interval, a certain cumulative density function with continuous and differentiable features should be incorporated to describe the relationship between response variable  $Z$  and its random characteristics. In a common linear model, the mean of the response variable is assumed to be linearly related to the predictor variables  $X$ . Since the mean implicitly depends on the stochastic behaviour of the response, and the predictor (explanatory) variables are assumed fixed, thus the aforementioned certain cumulative density function provides the link between the random (stochastic) component and the systematic (deterministic) component of the response variable  $Z$ . Hence, Nelder and Wedderburn [1972] refer to this function as a link function. In addition to the logistic function, an other frequently used link function is normit. The normit link function is

$$g(p) = F(p)$$

is the inverse of the cumulative standard normal distribution function, which is

$$F(x) = F(x) + (2\pi)^{-1/2} \int_{-\infty}^x \exp(-z^2/2) dz$$

In the literature the more familiar term Probit is often used. In financial failure prediction, Probit procedure is also employed to predict a bankruptcy event. Like the Logit model, Probit is one of conditional probability. It can estimate the probability of the occurrence of an outcome conditional on the predictor variables. Probit following the Logit procedure overcomes the boundary problem identified above by transforming the linear probability

model so that response variable will be bounded within the  $[0, 1]$  interval. Meanwhile it does not require restrictive assumptions such as multivariate normality and homogeneous dispersion across groups, and yields the meaningful conditional probability on a given observation, which has an insignificant difference from Logit's outcome in practice.

From these viewpoints the Logit and Probit models would appear to have a similar algorithm. However, there are some advantages in Logit procedure. The apparent merit is its simplicity in computation. Moreover, being asymptotic at the extremes, where a small change in the independent variables is unlikely to materially affect the outcome also makes the logistic function preferable. An other advantage of the Logit over Probit is that differences on the logistic scale are interpretable regardless of whether the data are sampled prospectively or retrospectively [McCullagh and Nelder, 1989]. Besides, there are some difficulties when the Probit model is extended to handle polytomous dependent variable situations. Aldrich and Nelson [1986] stated this point

"Unfortunately, while there are, again, an infinite number of such forms, they (other link functions) turn out to be infeasible to estimate. Multinomial Probit, for example, involves probability expression that are multiple integrals of the multivariate normal density. While accurate and simple approximations are available for the integral of the univariate density, comparable approximations are feasible for the multivariate integrals only up to about the fourth order. Beyond this dimension, computation is impractical. In other words, multinomial Probit, Gompit and the like logically possible but impractical." (p. 39).

On the basis of the above advantages and its greater popularity in the literature, Logit procedure is chosen as the representative of conditional probability methods in our comparative analysis.

## **2.8 Comparisons of MDA and Logit**

### **2.8.1 Theoretical Comparisons**

Classifying an observation into one of two (or several) populations is multivariate discriminant analysis. Relating qualitative variables to other variables through a logistic cdf functional form is logistic regression. These two methods, the most widely used statistical procedures in empirical studies of bankruptcy prediction models, are closely related particularly with respect to the classification problem of forecasting response for new

observations. In the classification problem,  $Y$  is assumed to be a discrete variable representing the different group, and  $X$  is a vector of predictive (explanatory) continuous variables. Discriminant analysis and Logit method can be the alternate techniques of characterising the joint distribution of  $(Y, X)$ . Discriminant analysis starts from the  $X$  variables conditional on  $Y$  with the assumptions of multivariate normal distribution of  $X|Y$  and equal variance-covariance matrices across the  $X$ . Logit method focuses on the distribution of  $Y$  conditional on the  $X$  which is assumed to be logistic. From another perspective, discriminant analysis can be thought of as treating the endogenous variables in Logit as the independent variables and asking, given  $Y$ , how the distribution of the  $X$ 's can best be described.

For purposes of prediction, Logit and MDA can be used interchangeably. However, there are some distinctions between these two methods

- (1) The Logit model does not share the assumptions of discriminant analysis. Given that the violation of these assumptions in realistic data is not unusual, it seems that the Logit model would be superior to discriminant analysis. However, it could sometimes be argued, the assumptions in MDA are unimportant if the only purpose of the model is to develop a discriminating device.
- (2) Logit is applicable for a wider range of distributions than is MDA. In other words, Logit procedure results from a wide variety of underlying assumptions about the explanatory variables [Anderson, 1972]. Thus, a logistic model is more robust than MDA. However, if the normality of  $X|Y$  can be satisfied, then discriminant analysis is the true maximum-likelihood estimator and therefore is asymptotically more efficient than the Logit maximum likelihood estimator (MLE) [Maddala, 1983].
- (3) Logit procedure allows for the independent variables to be discrete in contrast to MDA analysis. In empirical studies, failed and nonfailed firms were often matched according to criteria such as size and industry. The usage of these discrete or dummy variables obviously violates the normality assumption which is required in discriminant analysis but not in Logit technique. Although it is still unclear that the impact of the matching procedures is a gain or a loss, it would seem to be more fruitful to actually include these variables as predictors rather than to use them for a matching purpose. Consequently, the Logit method is favoured when macro or dummy variables are used in empirical studies.

- (4) Discriminant technique is basically a multivariate method that assign a score to each element in a sample using a linear combination of independent variables. The reduction of several financial dimensions of a problem to a single score is appealing. However, the questions about whether so many factors and dimensions of a complex financial problem like bankruptcy can validly be reduced to a single score, or whether crucial information would be lost during the process of reduction are also raised. In contrast, the Logit procedure does not reduce all dimensions of independent variables to a single cutoff score. Rather, it assesses each relevant independent variable and comes up with a probability of bankruptcy, given that a company belongs to a certain sample.
- (5) Logit procedure furnishes the meaningful probabilities of bankruptcy, and the model is thus relatively easy to interpret, while the output of MDA model is a score which has little intuitive interpretation. The score is basically an ordinal ranking device, although, if prior probabilities of the two groups are specified, then, MDA is able to derive posterior probabilities of failure. But, this Bayesian revision process will be invalid or lead to poor approximations unless the assumptions of normality and equal dispersion are satisfied [Ohlson, 1980].

### **2.8.2 Empirical Comparisons**

Many empirical studies have been conducted to compare the classification performances between Logit and MDA. Halperin, Blackwelder and Verter [1971] compared Logit and MDA for non-normal data. They reported that the estimators of coefficients from MDA can be quite biased for data consisting entirely of binary variables as well as a mixture of binary and continuous variables. Further work by O'Hara et al. [1982] and Hosmer et al. [1983a, 1983b] also showed that these estimators can be severely biased for mixed continuous and discrete variables. Press and Wilson [1978] indicated in a comparative study of Logit and discriminant analysis concerning breast cancer that Logit was mildly better than MDA for correct classification when applied to the holdout sample, regardless of using the different number of dummy variables which clearly violated MDA assumptions.

Crawley [1979] concluded that Logit is preferable to MDA when the group-conditional distributions are clearly non-normal or their dispersion matrices are clearly unequal.

Ohlson [1980] was the first to use Logit analysis to predict financial distress. Data for the model consisted of 105 failed firms and 2058 unmatched nonfailed firms from the year 1970-1976. All observations were used to derive the model and no holdout sample was used for validation. The overall misclassification rate was minimised at a cutoff point of 0.038. At this point 12.4% of bankrupt firms and 17.4% of the nonbankrupt firms were misclassified. However, if applied to a population composed equally of failing and nonfailing firms, Ohlson's model would have an expected overall error rate of 14.9%. These results appeared to be somewhat worse than those of previous studies. Ohlson provided four possible explanations for this. First, the lead time from the last fiscal year to the filing of bankruptcy is longer in his study than in previous studies. Second, the data he used was from the 1970s, which was later than for previous studies. Third, the selections of financial ratios are different. Finally, the choice of techniques may affect the results.

Collins and Green [1982] applied multivariate discriminant analysis, the linear probability model and logistic regression to test the prediction ability in a holdout sample of healthy and failed credit unions. They found that the Logit model was only modestly superior to MDA in overall classification accuracy. While classifying failed firms, Logit markedly surpassed the MDA. That is, Logit procedure can substantially reduce the Type I error rate. This outcome is important since misclassifying a bankrupt firm as a healthy firm is much more costly than misclassifying a healthy firm as a bankrupt firm. Zavgren [1985] built a Logit conditional probability model in predicting business failure. She assessed the previous studies made by Altman [1968], Deakin [1972], Edmister [1972], Wilcox [1971b, 1973], Blum [1974] and Diamond [1976], and found that some inappropriate estimates of coefficients occurred due to the loose assumptions of discriminant analysis. Strong evidence from Jones's [1987] research has also shown that Logit would provide more reliable classification accuracy than MDA. Hopwood, McKeown and Mutchler [1990] conducted a sensitivity study on bankruptcy models departure from normality in Probit, Logit and MDA techniques. They found that the Logit model was the most robust to non-normality.

However, Efron's [1975] study indicated that if the normality of independent variables is met, then the MDA is considerably more efficient than Logit. Further, Amemiya and Powell [1983] showed that for purposes of classification, MDA does quite well even if the independent variables are binary, which is clearly not a normal distribution. But this conclusion may be more likely to hold for discrete variables than for continuous ones. In Hamer's study [1983], a comparison of MDA and Logit yielded no statistically significant differences in overall accuracy. These mixed results indicate that no conclusive evidence has been reached on which method, MDA or Logit, produces better classification power.

In addition to modelling assumptions, MDA and Logit would not provide similar information for their output. In standard MDA, the result Z score is a composite value, and is compared to the cutoff point to decide which group the observation should belong to. This dichotomous partition of the outcome might be less useful for decision-makers such as capital stock investors, bond purchasers, accounting auditors and commercial loan makers. As Martin [1977] criticised, the user may be capable of varying levels of response to risk of failure if this meaningful information is given. Chesser [1974] discussed a noncompliance loan to emphasise this occurrence. He stated that a noncompliance loan means that the borrower would bargain with the lender to reach an agreement which can be less favourable to the lender than those specified in the original default agreement. In this situation the risk of failure is needed in order to make differential adjustment in the risk premium on interest rates and loan indentures. The Logit model, in contrast to MDA, can produce this valuable information. When the coefficients derived from Logit are applied to an individual in the sample, the resulting value measures the "vulnerability" to failure [Korobow and Stuhr, 1975], or the "propensity to fail" [Martin, 1977, p.257]. Some arguments may arise that the MDA also offers the probability of bankruptcy for each observation, yet under the assumptions that the obtained Z scores are of normal distribution. Martin [1977] posited that the probabilities generated by discriminant analysis most ordinarily used involve a subjective assessment of the probability associated with a particular discriminant score. He investigated various probabilities produced by a variant of discriminant function using a maximum likelihood estimation technique to assess probability. On the other hand, using a Logit model, he tested the results of this estimation



against the null hypothesis that the probability of failure is equal to the prior probability in the population according to same data. Martin found that discriminant analysis, regardless of linear or quadratic form had likelihood functions significantly lower than the null hypothesis. This would mean that the null hypothesis would provide a better probability estimate than either discriminant function. However, the Logit model had a likelihood function significantly higher than the null hypothesis. This indicates that the Logit procedure would provide a better probability estimate than that in MDA.

Furthermore, he pointed out that when the proportion of two groups in the sample is disproportionate with the proportion in the population, the probabilities obtained from the discriminant function may be very inaccurate, although the classification results improve by exaggerating the size of the smaller group. Thus, there existed large biases in predicted probabilities and prediction capability conclusions, since most analysis used equal-sized matched samples in the bankruptcy prediction model, which are strongly incompatible with the probability of bankruptcy in the population. Ohlson [1980] asserted that the obtained probabilities can not be relied upon if the assumptions of normality and equal group dispersion are not satisfied. This conclusion was consistent with the evidence shown in the Press and Wilson [1978] research. But Altman and Spivack [1983] found that the Z score rankings obtained from discriminant analysis were closely correlated with Standard and Poor's bond rating. In order to assess if the probability of MDA was unreliable, Hamer [1983] conducted an experiment by comparing the estimates obtained in discriminant analysis with those obtained using Logit. Four different data sets used previously for predicting bankruptcy were employed for each of five years before bankruptcy in these two techniques. Thus 20 MDA models were compared to 20 Logit models. Spearman rank correlation was applied to test the correlation of MDA rankings and Logit rankings in each of the 20 cases. The results revealed that these two techniques' probabilities were highly correlated and seemed to be comparable. Thus, the empirical evidence about the probability generated by discriminant analysis from these studies seems to be mixed.

## 2.9 Summary and Conclusions of MDA and Logit

This chapter discussed the traditional bankruptcy prediction methods—discriminant functions and conditional probability approaches. The most important ones are linear discriminant analysis and the logistic regression (Logit) model respectively.

Both linear discriminant analysis and the Logit method fit the data being studied in order to maximise the predictive power of the equation in the model. That is, the methods ensure that the sample correlation between the predicted and actual values of the response variable will be as large as possible. In discriminant analysis, the method tends to maximise the proportion of observations that are correctly classified in the sample, subject to the assumptions about normality in independent variables and variance-covariance homogeneity among the population. Logit procedure, one of the conditional probability models, determines its coefficients so as to maximise the joint probability of bankruptcy for the known bankrupt companies and the probability of nonbankruptcy for those companies that have not gone bankrupt. However, it does not share the demanding assumptions of linear discriminant analysis. Thus from the theoretical perspective, the advantages of the logistic approach to discrimination are, firstly, that it can be used with equal facility whether the variables are discrete or continuous and, secondly, that the estimation procedure is suitable under many different assumptions about the underlying distributions. Moreover, it yields estimates of posterior probabilities and likelihood ratios, which may be one of concerns of a study as in bankruptcy prediction. Nevertheless, if the normal distributions in independent variables are obtained or just slightly violated, then the discriminant analysis is considerably more efficient than the Logit procedure. On the other hand, from the empirical perspective, previous comparative studies of these two methods do not offer conclusive evidence in favour of either of them for purposes of classification accuracy. The choice of MDA or Logit method perhaps depends on the use of data sources, the nature of variables, the size of sample, or the intended results.

These two conventional statistical discriminating approaches will be compared in terms of their predictive abilities with those of newly developed artificial neural network techniques for both the simulation and empirical studies in this thesis.

## Chapter Three

# THE ARTIFICIAL NEURAL NETWORK

### 3.1 Introduction of Artificial Neural Networks

The human brain, the most complex computer device known to man, consists of substantial neurones. Each neurone is a simple microprocessing unit, which receives and combines signals from many other neurones. The neurone comprises nucleus, dendrites, soma, synapse and axon four parts. The main function of soma is handling the signal received from all directions of dendrites. If the combined signal is strong enough, it activates the firing of the neurone, which produces an output signal. The links between neurones are called synapse, which propagates the interneuron's signals. The tens of billions of neurones densely interconnected in the brain cause the ability to remember, think and problem solve. This has inspired many scientists to attempt a computer modelling of the brain's operation. One of the results has been the artificial neural network (ANN).

ANN tries to emulate the structure of the human neurones system and the operating way of processing information. The outcome may be knowledge representations based on massive parallel processing, fast retrieval of large amounts of information and the ability to recognise patterns based on experience. Paraphrasing Hecht-Nielsen, a pioneer in the development of neurocomputers, Caudill [1990] described an artificial neural network as

**"A computing system made up of a number of simple, highly interconnected processing elements, which processes information by its dynamic state response to external inputs"**

In effect, the early research on artificial neural networks was not successful until the mid 1980s. The renewed interest in ANN has recently been driven by two forces

- (1) The improvement in new hardware technology.
- (2) The development of the backpropagation algorithm [Rumelhart et al., 1986]

The first force involves the invention of parallel processing technology, which consists of a great number of independent processors operating at a very fast speed. Research suggests that the continual development of inexpensive microchips as well as the parallel processing concept offers the best architecture for ANN modelling and its future massive computation [Weems et al., 1991]; [Bharadwaj et al., 1992]; [Nolen, 1992].

The other driving force is the ANN algorithm renovation itself. Before the development of backpropagation neural network (BPNN), the field of neural network called Perceptron, which is the earliest neural network, experienced a lack of progress because some crucial shortcomings could not overcome. The Perceptron was invented by Rosenblatt [1959, 1962], one of the pioneers in the development of neural computing. His important "Perceptron" sparked a great amount of research interest in neural computing. In his very simple model, only a input and output layers are presented as illustrated in Figure 3.1.1. The input layer has several input units  $X_i$ 's which are linked to a single output unit  $Y$ . Input units first accept the information from the outside of the network and transmit the information into the output layer by summation of the multiplication of input units associated weights. If the sum of the weighted element is greater than a threshold (say 0.5), then the output is considered to be 1; otherwise, the output is deemed to be 0. For the simplest class of Perceptrons without any hidden layer (intermediated layer), Rosenblatt [1962] was able to prove the convergence of a learning algorithm, a way to change the weights interactively so that a desired computation was performed. However, the fatal limitation was presented by Minsky and Papert [1969]. They proved that the Perceptron without a hidden layer was unable to solve the Exclusive OR (XOR) function. The XOR problem requires a single output unit to be turned on (+1) if one or the other of the two inputs is on but not when neither or both inputs are on. This may be the result of the linear nature of the Perceptron. In other words, Perceptron can not cope with the nonlinear problem, even in the simple case such as XOR.

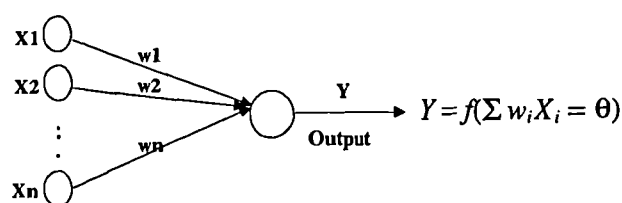


Figure 3.1.1 The Perceptron Concept

In 1986 Rumelhart, Hilton and William [1986] proposed a training algorithm for the layered machine called a generalised backpropagation algorithm. Backpropagation (BP) neural network improved at least two important points of the Perceptron by

1. Introducing the hidden layer between the input and output layers and allowing the interaction effect between input units
2. Replacing the original step transfer function with the differentiable transfer function and allowing the gradient steepest decent method to adjust the weights of network

The algorithm has given a new life to neural network and has been subsequently applied in many areas. Though it is not yet the perfect general algorithm capable of teaching an arbitrary computational task to a neural network, it can solve many problems which the simple two-layer Perceptron could not (such as XOR). As a result, much current research, including that of the classification problem, is centred on backpropagation and its extensions.

## **3.2 The Basic Structure of A Neural Network**

This section we will discuss the major elements of a neural network. An ANN model is composed of the following basic components: (1) Layer and Node, (2) Connection and Weight, (3) Processing Element (PE), (4) Network Operation

### **3.2.1 Layer and Node**

Layer consists of a set of processing units. There are three types of layers: input layer, hidden layer and output layer. The nodes present processing units in each layer. The node in the input layer accepts input values from outside of the system, the nodes in the hidden layer process the value from the input nodes and produce intermediate results, then transmit them to the output layer. The nodes in the output layer produce the output value based on the hidden node value and threshold status.

### **3.2.2 Connection and Weight**

The connection between two nodes is called link. It represents a flow of information. The network can be fully connected or partially connected. The fully connected network has each node in one layer connected to every node in the next layer. The partially connected network has nodes partially connected to some nodes in the next layer. Each connection has a corresponding weight which represents the stored knowledge of the network. The weights express the relative strength (or mathematical value) of the initial entering data or the various connections that transfer data from layer to layer. The signals from the input units to a processing element are modified by these weights prior to a processing element. In other words, weights express the relative importance of each input to a processing element. Weights are crucial; it is through repeated adjustments of weights that the network "learns".

### **3.2.3 Processing Element, PE**

Analogous to the biological neurone in the brain, "processing element (PE)" is the basic unit in an ANN. The internal activity is operated by two functions. The first is summation function. It receives the output value of other processing elements as its own input, multiplying this input values ( $X_s$ ) by the weights ( $W_s$ ) and totals them together for a weighted sum.

The other is activation function (transfer function). Activation function is needed to introduced nonlinearity into the network. Without nonlinearity, hidden units would not make nets more powerful than just plain Perceptron. The reason is that a composition of linear functions is again a linear function. The other purpose for activation function is that this transformation is to modify the output levels to a reasonable value (e.g. between 0 and 1, or -1 and 1). Without such transformation, the value of the output may be very large, especially when several layers are involved. Almost any nonlinear function does this job, but the sigmoid and tangent function are the most common choices.

### 3.2.4 Network Operation

Learning is a process of adapting behaviour in response to stimuli presented at input and output value to modify the connection weights and thresholds. That is, a network gradually learns the relationships between many input/output pairs by adjusting the strength of the connections or weights between processing units. When an input node is activated, it sends a signal or information to the nodes or processing elements in the intermediate hidden layer across weighted parallel connections. Each node in the hidden layer in turn sends information to the nodes in the output layer which provide a weighted output. Once the set of connection weights have been found for a particular pattern, the network can be trained to recognise the correct response when given another input pattern. There are three kinds of learning strategies

1. **Supervised Learning:** supervised learning involves outright comparison of the output of the network with known correct answers. A certain rule is provided to direct the modification of weights. The weights are adjusted by means of supervising the minimisation of cost function or error function. This behaviour is like that of a knowledgeable teacher who guides us in a correct direction. Supervised learning incorporates decisions about when to turn off the learning, how long and how often to present each input/output pair for training, and information about performance. In supervised learning, for each input stimulus, a desired output stimulus is presented to the system and the network gradually configures itself to achieve that desired input/output mapping.
2. **Unsupervised learning:** the network has only input value and no desired output is shown. Therefore, no definite rule can be learned. This case is also called self-organisation. In unsupervised learning only input stimuli are shown to the network and the network organises itself internally so that each hidden processing element responds strongly to a different set of input stimuli or closely related group of stimuli. These sets of input stimuli represent cluster in the input space, which typically represents a distinct real world concept.

This process has no external teacher and self-organises presented data and discovers its emergent collective properties.

3. Associative learning: this case falls between supervised and unsupervised learning. Although desired output is provided, it is incomplete instead of a definite desired output as in supervised learning. It must infer the correct status from incomplete information. This case is referred to as reinforcement learning where an external teacher indicates only dichotomous outcome in response to an input.

Whatever kind of learning rule is used, an essential characteristic of any network is its learning mechanism. It is the fundamental difference between the artificial neural network and conventional artificial intelligence.

### **3.2.5 Feedforward Network vs. Recurrent Network**

Feedforward network means that the role of each node, except in the output layer, is just to feed an input pattern from the lower layer to the higher layer. There are no connections leading from a unit to units in previous layers, nor to other units in the same layer, and nor to units more than one layer ahead. In other words, in such networks information flow is all in one direction. There are no feedback loops from a unit to a previous one. Every unit feeds only the units in the next layer. In contrast, networks that are not strictly feed-forward, but include direct or indirect loops of connections, are often referred to as a recurrent network.

## **3.3 The Difference between ANNs and ES**

Unlike a traditional expert system (ES), where professional knowledge must be made explicitly in the form of rule, neural networks generate their own rules by learning based on the example itself. An expert system depends on the representation of the expert's knowledge as a series of IF-THEN conditions or rules, known as knowledge base. These rules must first be determined by observing human experts, then programmed into the ES using special languages such as PROLOG or in shells such as Knowledge Craft, ART etc. These processes require enormous time and effort to extract the inferences upon which



the system is based. More importantly, an expert is unable to use inductive learning and inference to adapt the rule base to changing situations. In another way, once the system is functional, making even minor changes to the knowledge base can be a complex and expensive process because of the intricate relations between the rules forming the knowledge base [Coats, 1988]. However, dynamic situations are always the characteristics of financial and management environments. In contrast, neural networks do not require a predefined knowledge, and can automatically capture patterns among the given examples. When the situation changes, the neural network automatically responds to changes in the problem environment by adjusting weights. It has a self-maintaining function. Therefore, instead of the static nature of the expert system, a neural network is fundamentally dynamic. It continues to adapt and improve as it is exposed to new information.

Another problem with expert systems, as pointed out by Coats [1988], is that an "expert system cannot really deal with erroneous, inconsistent, or incomplete knowledge because most expert systems rely on rules that represent abstracted knowledge of the domain (i.e., the problem space) and thus the expert system are not able to reason from basic principle" (p.80). That is, expert system knowledge is fixed and predefined, and cannot tolerate minor component failure and employ common sense or make an "educated guess" to amend it. Thus, if the input information is incomplete, ambiguous or noisy, the expert system is unable to perform effectively. In this respect, neural network has its appeal. It can filter out noise and isolate useless or incomplete information in order to recognise patterns without impairing the entire system. Learning is achieved through a learning rule which adapts or changes the connection weights of networks in response to the input and the desired outputs in respect to those inputs.

However, the advantage of neural network is also linked to its disadvantage. The automatic "black box" learning process makes it difficult to trace the steps by which the output is reached. The optimal status represented by the matrix of connection weights can not be translated into clear information to the user. In other words, the neural network algorithm cannot tell the user how it processed the input information in order to reach a conclusion. On the other hand, an expert system can be broken down into discrete steps or series of operations. To some outsiders, the absence of a bright identifiable internal

logic is a severe obstacle to the acceptance of neural networks. Table 3.3.1 shows the advantages of neural computing over traditional artificial intelligence methods.

**Table 3.3.1 Comparison of Neural Network and Expert System**

<b>Neural Network</b>	<b>Expert System</b>
Example based	Rule based
Domain free	Domain specific
Finds rules	Needs rules
Little programming needed	Much programming needed
Easy to maintain	Difficult to maintain
Fault tolerant	Not fault tolerant
Needs (only) a database	Needs a human expert
Fussy logic	Rigid logic
Adaptive	Requires reprogramming

Source: [Samdani, 1990]

### **3.4 Backpropagation and Generalised Delta Rule (GDR)**

The backpropagation (BP) algorithm was invented independently several times, by Bryson and Ho [1969]; Werbos [1974]; Parker [1985]. But it did not attract much attention until Rumelhart et al. [1986] proposed this algorithm and established its important status. Now the BP is the central to much current work on learning process in neural networks.

A typical BP network always has an input layer, an output layer and at least one hidden layer. In the learning process, a set of training samples is needed, which includes input value and definite output value (it belongs to one of supervised learning). The units in the input layer do not perform weighting or nonlinear transformation. They simply sends its signal to each of the units in the hidden layer. The network initially gives a set of random numbers as start weights. According to the feedforwarding process, the information about the differences between computed and desired output is propagated backwards through the network and is used to update the connection weights. The basic updating rule in the learning process is gradient descent adjustment. Hence, this algorithm is sometimes

viewed as an enhanced version of the stochastic gradient-descent optimisation procedure [Berry and Trigueiros, 1993]. Details about the mathematical form can be found in Appendix I.

The process we have derived in BP is called generalised delta rule (GDR). The GDR contains basically the same idea as Perceptron, however, are more complex. The difference is in the way of calculating the delta. The delta in the Perceptron is simply the difference between the target value and the actual value (the system output), while the delta in the GDR is a function of the difference and the first derivative of the node. The GDR can be applied to any multi-layered system and thus is called the **generalised delta rule (GDR)**.

In summary, the GDR finds weights which minimise the sum of square errors between target values and system-produced values for all units in the output layer. Because the error function has a non-linear form with respect to the parameters (weights), the algorithm uses a gradient descent method to minimise the error function. The algorithm consists of two passes, a forward pass and a backward pass. In the forward pass, the system calculates the output value for each node based on input values and associated weights. The system then calculates the deltas from the highest-layer (output-layer) and propagates the errors or deltas backward to the lowest layer.

### **3.5 Variations on Standard Algorithm**

There are many parameters varying within the standard BP nets, including the learning rate, momentum, weights updating rule and different cost functions etc. On the other hand, the network architecture itself (number of layers, number of units for each layer) affects the performance substantially. In this section we focus mainly on the parameters of modifications, keeping others fixed. In the next section, a newly developed algorithm will be presented in order to remedy the drawbacks of the GDR algorithm. The question of optimal architecture will be discussed later.

### 3.5.1 Learning Rate

An issue related to the GDR is how to choose the learning rate. Because backpropagation is a gradient descent algorithm, it requires a scale factor that indicates how far to move in the direction of the gradient during the training process. This factor is called the learning rate and is denoted as  $\eta$ . That is, the learning rate  $\eta$  is a gauge of the rapidity of convergence every epoch. A too large or too small  $\eta$  will cause negative inferences to converge [Von Lehmen et al., 1988]. If  $\eta$  is too small, the learning process can be very slow. On the other hand, a too large  $\eta$ , although approximating the minimum of error function by large weights updating, can oscillate widely or may cause a move through weight space that "overshoots" the optimal solution [Knight, 1990]. The problem essentially comes from cost-surface valleys with steep sides but a shallow slope along the valley floor. One of solutions to cope with this problem is to adjust this value during the learning process. In effect, "tweaking" of the learning coefficient is necessary in practical applications. A large learning constant is needed when the current weight vector still has a substantial distance to travel to the optimal position. As the weight vector approaches the point at which error function is minimised, smaller and smaller step may be desirable. Another approach, adding a momentum term which is based on the similar concept [Plaut et al., 1986], has been more commonly used and is shown effective.

### 3.5.2 Momentum

Momentum is a weighted term which considers a previous change in the weight when calculating the present change. The concept of momentum term is introduced to resolve the dichotomous dilemma involving minimum learning rate. It will stabilise the convergence with error surfaces containing long ravines that display sharp curvatures across the raven and a gently sloping floor [Wasseman and Schwarts, 1987]. The idea is to enter some inertia  $\alpha$ , so that the previous delta weight is fed through to the current delta weight.

The effect of the momentum tends to reinforce a general trends in the delta weight term whereas oscillatory behaviour cancels itself out. Then the effective learning rate can be made larger without divergent oscillation occurring .

### 3.5.3 Cumulative Update of Weight

Another technique that can speed up the convergence is to only update the weight in a batch way rather than entering every individual training sample. This is referred to as cumulative backpropagation because the delta weights are accumulated until the complete set of input/output pairs is presented. The idea is expressed as

$$\Delta W = \sum_m \Delta W_{ij}^m / \sqrt{N} \quad \begin{array}{l} \Delta W_{ij}^m = \text{the change in weight of the } m\text{th training sample} \\ N = \text{sample size of training set} \end{array}$$

If the epoch is not too large, the above approach can lead to faster convergence since an individual update only reduces the error function for a particular pair while probably increasing other component error functions, whereas the global update will always reduce the overall error function. But if the epoch is large, the benefit of using an overall error function may be lost because of more calculations involving this cumulative updating.

### 3.5.4 Alternative Error Function

The error function defined in Appendix I is proportional to the square of the Euclidean distance between the desired output and the actual output. This is not the only possible choice. Any other differentiable function that is minimised by  $T_j = A_j$  ( $T_j$  = the target pattern,  $A_j$  = the actual output pattern) can derive a corresponding update rule. One of the choices proposed [Baum and Wilczek, 1988]; [Hopfield, 1987]; [Solla et al., 1988] has received particular attention

$$E = \sum_j \left[ \frac{1}{2} (1 + T_j) \log \frac{1+T_j}{1+A_j} + \frac{1}{2} (1 - T_j) \log \frac{1-T_j}{1-A_j} \right] \quad (3.5.1)$$

Like the quadratic cost function, (3.5.1) is always positive except when  $A_j = T_j$  for all  $j$  where  $E=0$ . However, it has a natural interpretation in terms of learning the correct probabilities of output nodes  $j$ , if  $1/2(1+A_j)$  is used for the probability that the hypotheses represented by  $j$  is true. When  $A_j = -1$  means definitely false, and when  $A_j = +1$  means definitely true. Similarly,  $1/2(1+T_j)$  is interpreted as the target set of probability. Then

information theory suggests the relative entropy (3.5.1) of these probability distributions as a natural measure of the difference between them [Kullback, 1959]. Its advantage is also shown qualitatively since it diverges if the output of one node saturates at the wrong extreme, instead of just approaching a constant in quadratic measure, which can lead to the learning hang around on a relatively flat plateau of  $E$  for a long time. Given its nature, the cost function (3.5.1) has been shown to solve some problems that cannot be solved using the quadratic form, and is thus suitably applied in prediction problems when the training data are actually probabilistic or fuzzy [Hertz et al., 1991].

### 3.5.5 Different Activation Function

In BP learning, any differentiable can be used as the transfer function. Nevertheless, a linear transfer function gains no additional advantage in using the hidden layer. The hyperbolic tangent function is one of them. It is, in effect, quite similar to the sigmoidal function in shape (S-shaped). The sigmoid is a smooth version of a  $[0, 1]$  step function, whereas the hyperbolic tangent is a smooth version of a  $[-1, 1]$  step function. Because the output of the transfer function is used as a multiplier in updating the weights, a range of 0 to 1 means a smaller multiplier when the summation is a low value, and a higher multiplier for higher summations. This could lead to a bias towards learning higher desired outputs. In contrast, the hyperbolic tangent gives equal weight to low and high end values. However, the use of the sigmoid function has its appeal. First, applying the sigmoid function, the deltas will become the following simple form

$$\begin{aligned}\delta_j^{[n]} &= -(T_j - A_j^{[n]}) f(Z_j^{[n]}) \\ &= (T_j - A_j^{[n]}) A (1-A) && \text{if in the output layer} \\ \delta_\phi^{[v]} &= [-\sum_k \delta_k^{[v+1]} W_{\phi k}] \phi'(Z_\phi^{[v]}) \\ &= [-\sum_k \delta_k^{[v+1]} W_{\phi k}] A (1-A) && \text{otherwise}\end{aligned}$$

Another advantage of the sigmoid function is that it generates the continuous values between 0 and 1 which are suitable for the probability issue, rather than binary values used by the early neural network model. It gives more flexibility to the system.

Researchers have proved that a neural network with sigmoid function units can essentially fit any function and its derivative [Cybenko, 1989]; [Funahashi, 1989]; [White, 1989]; [Barron, 1992]. This is probably the reason why the sigmoid transfer function is most commonly used in ANNs.

## **3.6 Problems with GDR Backpropagation**

### **3.6.1 The Drawback on Long Learning Time**

There are some problems with the GDR backpropagation. The first is the speed of learning. A long learning time has been considered a major obstacle for its applications to real world problems. The GDR learning algorithm must spend much computational time for the weights to converge, and it scales poorly as tasks become larger and more complex. Although the parallel processing hardware has been developed to try to overcome this problem, for a complicated input space it still take a long training time. Moreover, in most cases this algorithm has actually been implemented just by software emulation instead of hardware support on the machine. Several techniques have been studied to speed up the GDR BP, such as the introduction of a momentum in the previously discussed learning rule [Rumelhart, Hinton, and Williams, 1986], the use of an alternative cost function instead of the standard quadratic error function [Fahlman, 1988]; [Krogh, Thorbergsson, and Hertz, 1989]; [Solla, Levin, and Fleisher, 1988], the dynamic adaptation of the learning parameters [Cater, 1987]; [Jacobs, 1988]; [Jutten, Guerin and Nguyen Thi, 1991], the application of a more elaborate search method [Becker and Le Cun, 1989]; [Parker, 1987] and the incorporation of a probabilistic learning algorithm [Specht, 1990]. Many of these either involve the variations of Newton's method, requiring the computation or approximation of second partial derivatives, or use the approximated higher-order derivative of the error function which provides more information about the shape of the weight space. However, these methods, in spite of reducing the rate of convergence dramatically, increase computation load and tend to not scale up very well as the problem size increases. On the other hand, the probabilistic

backpropagation, has the advantage of its simple computation and easy programming, but thus, will be offset by the sacrifice of classification accuracy.

### **3.6.2 The Drawback on Local Minima Solutions**

Another major problem encountered with BP is that of being trapped in the local minimum of the error function. Since the GDR algorithm tries to find the best weights and thresholds by merely using the first order condition minimising technique, it can not guarantee a global solution. The aim of minimisation of a multi-dimension nonlinear is to find the one position on the surface of the function where the value is at its global minimum. The negative of the gradient of the surface is expected to point towards the global minimum. However, there may be local minima which are holes in the surface that are indistinguishable from the global minimisation. If the global minimum is not known, there is no way to be certain that a minimum in the surface is a local minimum or a global one. In particular, in the presence of a local minimum, the BP sometimes fails to find a set of weights and thresholds for a network.

The existence of local minimum training errors in some nonlinear feedforward ANNs has been identified by several investigators [Sontag and Sussmann, 1989]; [Brady, Raghavan and Slawny, 1989]; [Blum, 1989]; [Gori and Tesi, 1992]. Dorsey, Johnson and Mayer [1994] also confirmed that the error surface for the ANN is frequently characterised by a large number of local optima. The most important factor causing local minimum is the inappropriate and aimless randomisation of the initial weights and thresholds. Lee et al. [1991] have indicated that the initial weight has a direct effect on the training speed and convergence to local minima. Kolen et al. [1990] have also demonstrated that the backpropagation is very sensitive to the initial weights. In a GDR backpropagation neural network, initial weights and thresholds are usually chosen randomly. That situation is just like a person who is searching for the top of Mt. Everest (a global minimum), and is dropped by parachute somewhere over Asia by a pilot who has lost the map. The person may find the highest mountain top within his field of vision, and may assume that it is the goal. If the top of the mountain he reached is just a little lower than Mt. Everest (a good local minimum), it would be satisfactory. On the other hand, the tip of a small hill would not be acceptable.



Researchers have tried several different approaches to alleviate the local minima problem. Rumelhart et al. [1986] designed the heuristic addition of the momentum to speed up the training and to push the movement from local minima. Fahlman [1989] introduced a term in the calculation of the gradient to ensure that the solution cannot go to zero, and thus never comes to a halt on flat surfaces. An other simple idea is to use incremental updating, choosing the patterns in a random order from the training set so that the average over patterns is avoided and the random order generates noise. Alternatively, adding noise explicitly by random, changing the weights slightly [Von Lehman et al., 1988] or adding noise to the training set inputs independently at each epoch [Sietsma and Dow, 1989] are also possible solutions. In each case, there seems no overall competitive approach. It may help a little to prevent the local minima, but may hurt or slow down the learning process considerably.

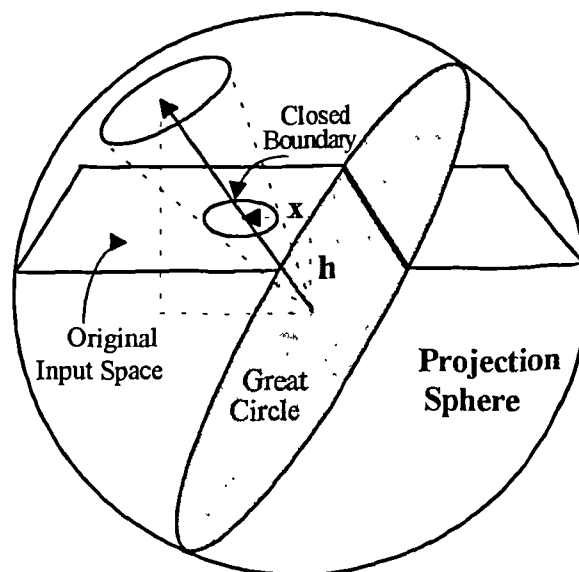
### **3.7 The Projection Neural Network**

The shortcomings of the GDR backpropagation network were presented in the preceding section. Alternative learning algorithms for solving these problems were proposed. However, they either solved just one of the problems at the expense of another, or lacked computational efficiency.

A powerful approach, Projection algorithm, was newly developed by Wilensky and Manukian [1992]. It overcomes all drawbacks of GDR backpropagation simultaneously. BPNN, although suffering from a slow training time and the potential to get stuck at local error minima, nevertheless, it offers the advantage of ensuring minimisation when it does converge to a solution. On the other hand, there exist other classification algorithms which train quickly but do not guarantee minimisation of the classification error. These networks include RCE (reduced Coulomb energy [Reilly et al., 1982], ART (adaptive resonance theory) [Carpenter and Grossberg, 1987], and the Cohonen type networks. They achieve fast training by placing prototypes with closed boundaries such as hypersphere at input data, match all input data to these prototypes by adjusting the radii, and thus update the positions and sizes of training points. The idea of the Projection is to

combine the advantages of these two classes of networks into a single one so that it can immediately place closed decision boundary prototypes around the input points and minimise the output error through gradient descent.

The BP algorithm partitions the input space into regions bounded by hyperplanes and curved surfaces through the nonlinear features of sigmoidal function in hidden layer. It starts its weights from open prototypes, then updates weights by adding the number of hyperplanes based on the error minimisation rule. However, linear algebra taught us that forming a closed region in  $N$  dimensions requires at least  $N+1$  hyperplanes. The more complicated the classification boundaries, the more regions and thus the more hyperplane or curved surfaces will be needed. For large  $N$ , this process can cause excessive training time. In contrast, the Projection algorithm initialises the weights and thresholds in  $(N+1)$  dimension space as a prototype with a closed or open boundary, providing a good starting point so that the network output is already close to a desired minimum. Then the network subsequently is trained as a GDR backpropagation network to further reduce the output error. Wilensky and Manukian [1992] have demonstrated that the Projection algorithm results in orders of magnitude reduction in training time and avoids a local minimum occurring. The theory of Projection network is presented as follows. For convenience of explanation, a 2-D inputs projected onto a 3-D sphere is illustrated



**Figure 3.7.1 The Projection Transformation and Formation of Boundary Surface**

The plane represents the original input space. We project an input vector  $\mathbf{X}$  in this 2-D plane onto a 3-D to get a new input vector  $\mathbf{X}'$  which is confined to lie on the 3-D sphere of radius  $R$ . Using the Pythagorean theorem, the components of  $\mathbf{X}'$  can be easily derived, and this can be expressed mathematically

$$\mathbf{X}' = R \left( \frac{h}{\sqrt{h^2 + \mathbf{X}^2}}, \frac{\mathbf{X}}{\sqrt{h^2 + \mathbf{X}^2}} \right)$$

$h$  is the distance between the origin of the plane (2-D) and the space (3-D)

$R$  is the radius of the sphere

The first term of  $\mathbf{X}'$  is the components of the original  $N$ -dimensional space. The second term is the component of the projected vector along the extra dimension. Note that if the decision boundary is a circle on the conic section of the sphere, its projection back onto the plane is a circle or an ellipse. If the decision boundary passes through the north pole, then its projection back onto the plane is a line. As a matter of fact, the projection of a circle from the sphere back onto the plane could be a circle, ellipse, hyperbola or lines on the plane, depending on the size and location of the circle. In other words, a large number of complex input space in an original hyperplane will be easily projected as a hypersphere on the  $(N+1)$  dimension. If  $R$  are chosen in an appropriate way, the inputs will project onto a good portion of the hypersphere and can be easily separated for classification. Therefore the weights and thresholds will begin with a good solution to the desired output since they must also lie on an  $(N+1)$  dimension hypersphere with  $|\mathbf{w}'| = R$ . Briefly, the trick of the unification of two types of networks is to project the  $N$ -dimensional input vector onto an  $(N+1)$ -dimensional hypersphere. The hyperplanes, provided by the GDR algorithm as decision boundaries to partition the input space, will then intersect with the hypersphere and thus the hyperspherical (closed or open) boundaries can be produced. Wilensky and Manukian [1992] further explained

"The reason for adding the extra dimension to the inputs before normalisation is to preserve all the information contained in the input vector, particularly its overall magnitude. In contrast, a simple normalisation of  $\mathbf{X}$  would confine the inputs to a hypersphere, but it would lose potentially valuable information contained in the magnitude of each input. This can be important if the radial direction contains important discriminatory information. In addition, such a scheme would not allow sufficient flexibility in the choice of the shape a prototype's decision surface." (p.360).

To construct a Projection network, the projected inputs  $X'$  serve as the inputs into a standard feedforward network with an additional node in the input layer. Thus the projected network has the option to choose and the ability to use either hyperplanes or hyperspherical prototypes or both for the hidden nodes in the solution of any particular problem, whereas, standard BPNN uses only hyperplanes and RCE uses only hypersphere.

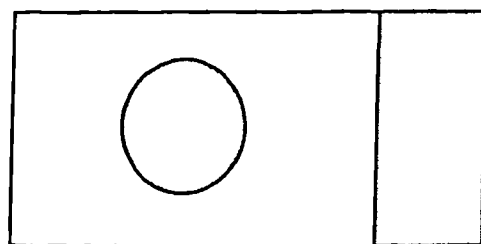
The advantages of using projection algorithm can be summarised

- 1. It combines the speed of a hyperspherical network with the error minimisation of GDR backpropagation.**
- 2. It partitions the input space into the region, which can easily put the inputs into different groups and, as a consequence, leads to more classification accuracy.**
- 3. It avoids the minimum which GDR backpropagation network often experiences by establishing a good starting point.**
- 4. It speeds up the training time by properly initialising the network weights and thresholds to input prototype.**
- 5. It has modular ability, allowing the flexible contribution of two or more networks.**
- 6. It reduces the number of required nodes in the hidden layer and leads to its more efficient use.**

To demonstrate clearly the concept and the advantages of the Projection neural network, some examples are presented:

#### Example 1

Consider the following circle and rectangle classification problem



The input space is divided into two separate regions: the grey and the white. To separate these points into distinct classes, backpropagation sets lines (2-dimensional hyperplanes) as boundaries between the grey and white regions. It starts a single line (one node in the hidden layer), then brings in another line for minimising the output error. In the end, it engages four lines (four hidden layer nodes) and adjusts their position until a good solution is reached. It takes 100,000 epochs to separate the grey and the white. On the other hand, the Projection network, using only two nodes in the hidden layer, begins by placing two circular prototypes with some reasonable radius. Then, it adjusts the position and radius of the prototypes until the class boundaries are matched as best as possible. Because the Projection network can use either circles or lines (hyperspheres or hyperplanes), it only needs two hidden layer nodes to solve the problem, as compared to four needed by GDR in this example. Furthermore, it takes just one-tenth learning time of GDR backpropagation to get an even better solution [Neural Computing, 1993]. Figure 3.7.2 displays the comparison of training process between the GDR and Projection algorithms on this problem.

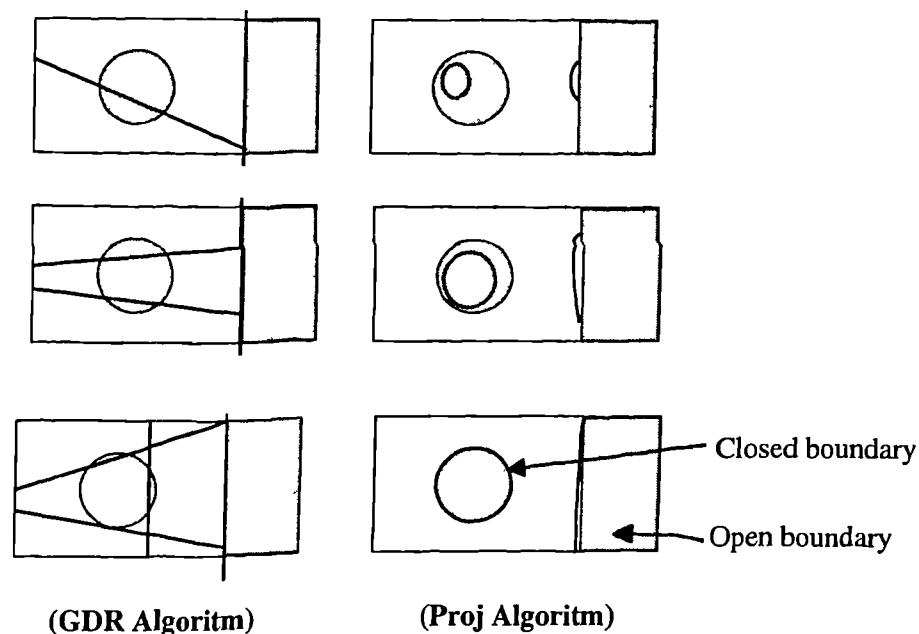
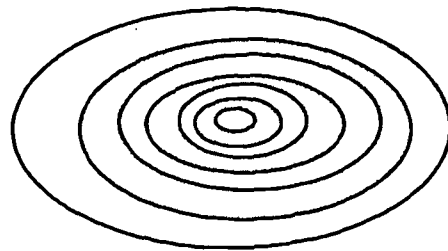


Figure 3.7.2 The Comparison of Training Process between GDR and Projection Circle and Rectangle Classification Problem

### Example 2

A well-known spiral problem consisting of several interlocked spirals is illustrated



The spiral problem was used as a benchmark comparison to test the performance of a network on a problem with a complicated boundary between classes. A GDR backpropagation network with one hidden layer is trapped in a local minimum and cannot find a solution. Lang and Witbrock [1988] employed the GDR backpropagation approach with two hidden layers of five nodes each. Nodes in the first hidden layer divide the input space into two regions along various angles. Nodes in the second layer subsequently employ combinations of these first layer features to produce curved patterns. Although this approach obtains some significant successive results of the spiral problem, it takes a long training time. In addition, the number of hidden layers remains a trial and error problem. By contrast, a Projection algorithm solves this spiral benchmark problem in only 600 training epochs by efficiently partitioning input space into closed prototype as well as the use of hidden layer effectively.

## **3.8 The Importance of Optimal Network Architecture**

The construction of architecture of a neural network is very important. An optimised architecture network can not only reduce computational training time but also improve generalisation ability. Therefore, it is desirable to find a technique to optimise the weights and thresholds for a given architecture as well as to optimise the architecture itself. Optimal architecture building involves determining the number of hidden layers and the number of nodes in hidden layer. Since it has been proven that a multilayer NN with one hidden layer can approximate any relationship between input and output

variables using an appropriate number of units in hidden layer [Cybenko, 1989]; [Hornik et al., 1990]; [Hornik, 1991], this provides a theoretical justification for the number of hidden layers. However, there is not yet a rule to decide the number of hidden nodes.

The number of hidden units dictates the space separability and thus the discriminating capability. That is, the hidden units in neural networks perform significant nonlinear data transformation for output units in order to produce arbitrary output functions. An insufficient number of hidden units may result in the network's inability to solve the problem it is supposed to tackle, because there are not enough parameters to model complex decision boundaries. On the other hand, if too many hidden units are employed, it may lead to spurious decision boundaries or may cause poor interpolation and unnecessary slow-speed convergence. Sietsma and Dow [1991] have demonstrated that when an excessive number of hidden nodes is used, the system will result in an overfitting problem. That means the system may pay undue attention to insignificant details and noise on the learning, inhibiting its capacity for generalisation on the prediction. Therefore, it is necessary to find a structure which has a minimum number of nodes in hidden layer and without a loss of generalisation capacity.

### **3.9 Past Studies on Determining Optimal Architecture**

#### **3.9.1 Trial and Error Approach**

Hirose et al. [1991] developed a process to determine the appropriate number of hidden units. Their original objective was to find a solution to avoid the local minimum problem. Basically, their algorithm is a kind of trial and error approach, including three steps: (1) check the total error every 100 epochs. If the error does not increase by 1% over the previous value, a new hidden unit will be added; otherwise the training continues for another 100 epochs. (2) the new set of weights should be randomly assigned or set to 0 when a new hidden unit is added. (3) Once the network converges, a hidden unit is removed and the network is trained again. This process is repeated until the network no longer converges.

The other method viewed as one of trial and error has been developed by Fogel [1990]. He claimed that the problem of choosing the optimal number of nodes and layers is analogous to choosing an optimal subset of regression variables in statistical model building. Fogel indicated that the output of a final-layer node before passing transfer function will asymptotically have a normal distribution if the previous layer has already a sufficient number of nodes. On the basis of this attribute, he derived the density functions of the residuals and the joint likelihood functions of two classes. The results will be compared to a measure called final information statistic (FIS), which is similar to the Akaike's [1974] information criterion (AIC). The FIS are computed for each proposed model to select the optimal number of hidden nodes. The model which has the smallest FIS value will be chosen. In this algorithm the proposed number of hidden nodes should be decided first, which involves a lot of trial and error processes. Therefore the technique requires many experiments and tedious computations.

### **3.9.2 Genetic Algorithm**

The idea of a genetic algorithm was derived from the biological system which can easily adapt to changing environments. Thus, some biological mechanisms, including reproduction, mutation and crossover, have been suggested as offering an approach to problem solving in various areas [Goldberg, 1989]; [Nygard et al., 1992]. Caudill [1991] proposed applying genetic algorithm principles, such as those in biological system, to evolve a neural network.

The reproduction operator copies a network's genetic description from one generation to the next. The mutation operator allows small, random change to occur in the genetic code, and thus usually corresponds to copying errors. The crossover operator involves two networks to combine to produce "offspring" whose genetic description is a combination of its "parents". The algorithm starts by generating an initial population randomly, the size of which is chosen by the user. When the initial population is generated, each member of the population is evaluated by an evaluation function. Two solutions of the population are selected and altered by crossover and mutation operators. Then they are combined to make offspring based on their fitness. The offspring are



inserted into the population. This second generation is trained and tested for their relative fitness, and is subject to the genetic algorithm to produce generation 3. This cycle is repeated until some stopping criteria are met. The stopping criteria can be the number of iterations or the level of fitness or convergence. In short, genetic algorithm maintains a population of promising solutions. Those solutions which do not fit well are supposed to die off. This mechanism is just like the survival of the fittest in biology. The new solution based on previous solutions will induce the search for a better solution to the problem. Finally, the search will arrive at an optimal solution. This methodology is obviously superior to the manual trial and error methods. Moreover the evolutionary nature of the algorithm enables the search for good configurations to proceed in a parallel, thus reducing the possibility of trapping in local optimal configuration. Nevertheless, maintaining a population as an evolutionary network requires large computer memory and system resources. Due to this drawback, the genetic algorithm searching technique seems unlikely to be practical especially for large network applications.

### 3.9.3 Weight Decay Technique

Hanson and Pratt [1990] developed a promising and effective algorithm called weight decay technique to build an optimal network structure. The idea behind this approach is that the network preferentially removes less useful connection weights. A mechanism causes the weight  $W_{ij}$  to decay to zero as the unit receives insufficient reinforcement and appears to be unnecessary. The simplest method to update  $W_{ij}$  value is  $W_{ij}^* = (1-\epsilon) W_{ij}$  (Where  $\epsilon$  is an assigned small positive number [Hanson and Pratt, 1990]).

The weight decay method is easy to implement. Under this technique the architecture is simplified gradually as unnecessary units and connections are removed from the network. However, a dilemma will arise in using this method. Because of the slow increase in connection weights, despite a small  $\epsilon$ , almost all of the weights decay to zero only offer a few times of updating  $W_{ij}$ . The vectors of hidden unit's weights become equal with further training, and thus the network degenerates to a single hidden unit structure which can rarely be trained successfully in most tasks. This seems to suggest that the weight decay method is more suitable for use only after the network has been trained.

### 3.9.4 Pruning Technique

One method similar to the idea of weight decay method is the pruning technique as introduced by Sietsma and Dow [1991]. Pruning refers to a process that examines a network, determines unnecessary units to the solution, and removes them from a network. In this process two stages are involved. In the first stage the noncontributing units are removed. The noncontributing units are (1) those which always produce a constant value across the training set, (2) those which always produce the same output as another unit, (3) those which convey the same information and differ only in direction across the training set. The second stage involves the elimination of units that are independent of the other units in the layer and give unnecessary information to the next layer. Sietsma and Dow [1991] demonstrated an example concerning the removal of unnecessary information units. The relevant layer with three units initially produces four patterns after the training process and may give approximate outputs as follows

Pattern	unit 1	unit 2	unit 3
A	1	1	1
B	1	0	1
C	0	0	1
D	1	1	0

Suppose that the next layer produces one output for pattern A and B, and a different output for pattern C and D respectively. Since the units 1, 2 and 3 are linearly independent, units 1 and 2 alone have already provided sufficient information to give a unique identification for each class to the next layer. Thus unit 3 is unnecessary and could be removed to minimise the network structure without loss any useful information. However, in practice, there is no rule to determine the noncontributing unit. Noncontributing units must be identified via manual inspection. For a large and complex problem, manual inspection is not feasible.

Mozer and Smolensky [1989] presented another way to remove the unnecessary units in the network itself. The method is a skeletonisation technique that trims the "fat" from a network through a relevance assessment. The relevance of unit  $i$ ,  $\mu_i$ , is denoted as

$$\mu_i = E_{\text{without unit } i} - E_{\text{with unit } i}$$

where  $E$  is the error of the network in training process. Based on the approximation of  $\mu_i$ , instead of the difficulty of calculating this value directly, the least relevant unit is identified and removed from the structure. This skeletonisation technique provides not only the way to trim the redundant units in the hidden layer but also offers a solution to find the contribution of each variable in input layer.

Based on the idea behind this algorithm, it seems to imply that when a weight is small, the impact of removing the connection from the network on the performance of the network is also small. Therefore the connection with small weight can be eliminated without impairing the classification accuracy. It also establishes a foundation to determine the effect of each independent variable on the output in multiple-layered neural network models, and thus the concept may help to solve the limitation of the interpretation of the significance estimation of input variables in ANNs (this will be discussed in a later section). Although this process, in most cases, has been tested to reveal good performance, the complicated computation is still an obstacle to implementation in practice. A simple method is expected to be developed to solve this problem. Principal Component Analysis (PCA) is perhaps one of these solutions.

### **3.9.5 Principal Component Analysis (PCA)**

Principal Component Analysis (PCA) is a technique that selects the most relevant variables from a pool of candidate variables. It is commonly used in multivariate analysis to reduce the dimensionality of the explanatory variables. PCA examines the covariance matrix of the variables and finds a few components which can explain the original variables very well on condition that the components are unrelated to each other. This idea can be employed in building optimal architecture in neural network to check whether there is any redundant information on the outputs of the hidden nodes, and then this can be removed from the network. The method follows the PCA technique used in multivariate analysis to inspect the covariance matrix of the outputs of the hidden nodes to test if the network has redundant units in hidden layer. Since each node in the hidden layer has as input the linear sum of the input variables and produces as output the

sigmoidal transformation of the input, if there is any redundant information, a hidden node can be represented by another hidden node or a set of hidden nodes. Therefore, the rank of the covariance matrix will be less than the number of hidden nodes. In other words, the principal components extracted by this process can be considered as the equations to define a hyper region, and its number is the number of equations required to define the region. This concept is similar to the example demonstrated by Sietsma and Dow [1991]. But Sietsma and Dow failed to build the mechanism into the algorithm to identify the unnecessary hidden node. The PCA method used to find the structure of the network can be expressed [Park, 1993]

- (1) Arbitrarily choose the appropriate large number of nodes. The maximum number required to perform the classification can follow Kolmogorov's theory [1968]. The number of hidden nodes in the initial network is assumed to be  $k$ .
- (2) Train the network, and all weights and thresholds are randomly assigned in this initial network.
- (3) Find the covariance matrix of outputs of the hidden nodes, that is  $k \times k$  matrix and apply the principal component analysis technique to the covariance matrix.
- (4) Obtain the eigenvalues of the matrix and identify the number of eigenvalues whose value is greater than 1, say  $k'$
- (5) If  $k'$  is less than  $k$ , choose these  $k'$  nodes out of  $k$  by examining the correlation between the hidden nodes and the selected principal component. Otherwise, it means that there are no redundant units in the hidden layer.
- (6) Retrain the network with a new structure, either by randomly reassigning the initial weights or using the estimated weights generated in selected nodes from step 5.

This process, unlike other methods which require many experiments with different structures, only trains the network for two different structures. Another advantage is that it may not require new learning when the principal components are highly correlated with the selected hidden nodes [Park, 1993]. Besides, the PCA is already a well-established technique in multivariate statistics technique and needs no assumption on the distribution of estimators.

### **3.9.6 Cascade Correlation Approach**

Cascade correlation is a learning paradigm invented by Fahlman and Lebiere [1990]. This algorithm can automatically determine its own optimal size. It starts with no hidden unit. The only connections are direct connections from the input layer to the output layer. The system then incrementally creates hidden units one at a time in order to predict the current remaining output error in the network. In contrast to ordinary feedforward design, each hidden unit of a Cascade learning receives input from all previous hidden units as well as from the input unit, just like cascaded connections. When a new hidden unit is created, this untrained hidden unit is referred as a candidate unit but is not yet permanently connected to the network. This new hidden unit is trained so as to maximise a measure of the correlation between its output and residual error at the output for the current training vector. If it cancels a portion of the forecast error, the unit is installed permanently into the network. This cycle of adding hidden units one by one is repeated until further addition of the hidden unit no longer shows any improvement or until the error is within a limit of toleration set by the user. Cascade learning has been shown to have several advantages. In addition to self-determining the configuration of the network, it also learns quickly and retains the structure it has built even if the training set changes [Fahlman and Lebiere, 1990]. In this study the Cascade learning paradigm will be applied both for the simulation and the empirical study in order to determine the appropriate number of hidden units when the ANN techniques are used.

### **3.10 The Significance of Input Variables**

The hidden layer provides the ability to explain the behaviour from the input layer to the output layer, and it also builds a obstacle to the interpretation of the impact of the input variables on the outcome. Seeing that an input unit is indirectly connected to an output unit through hidden units, the task of separating the contribution of each input variable to the output value is very difficult. Therefore the interpretation of weight values is sometimes one of limitations in the multilayer neural network model. Several researchers have attempted to examine the network dynamics to obtain the estimate of the degree of

the impact that an input has on the relative outperformance. The similar techniques used in the studies by Gorman and Sejnowski, [1988]; Klimisaukas et al. [1989]; Sen et al. [1992] and Refenes et al. [1994] to interpret the relative significance of the input variable attempt to find the change in output (Y) relative to the change in an input (X), namely the partial derivative of output to the input  $\partial Y/\partial X$ . For multilayer networks, the partial derivatives  $\partial Y/\partial X$  can be computed by applying the chain rule for derivatives repeatedly through the path g that connects the output node Y to the input node X. The final result of the  $\partial Y/\partial X$  for all such paths g linking the output node Y and the input variable X is obtained in the following way [Refenes et al., 1994]

$$\frac{\partial Y}{\partial X} \Big|_G = \sum_h W_{gh} K_h$$

$W_{gh}$  = the weight between node g in the input layer and node h in the hidden layer

$K_h$  is a function independent of any parameters in the input layer

$K_h$ s are the same for all input variables, because all changes reflected from the output layer through the hidden layers pass through the same paths for all input variables. As a consequence, the results suggest that a relative change in Y with respect to a change in input variables is affected primarily by the weights in the input layer. For instance, if the absolute value of the weights in the particular input variable are high, it can be concluded that change in this input will result in sensitive change in the output value. But the magnitude of this change can not be determined unless the value of  $K_h$  is obtained. However, the  $K_h$  value not only depends on the weights in the hidden layer, but also depends on the input layer. It can not be easily computed. Sen et al. [1992] developed an estimate method for  $K_h$  by combining all the weights and inputs. This approach was based on a weighted mean and can be used to obtain a better estimate of  $\partial Y/\partial X$ .

A simpler way was introduced by Baba et al. [1990]. They proposed a linear approximation method to extract the relative strength between each input and each output. This assumes that the input variables are not highly correlated, and offers good estimators of the coefficients of each input variable only at the point where other input

variables are fixed. The other approach to profiling the characteristics of each input variable is introduced by Yoon et al. [1993]. In his interpretation method, the strength of the relationship between an input X and an output Y was measured by

$$RS_{ji} = \frac{(W_{ki} * U_{jk})}{\sum_{i=0}^m ABS [\sum (W_{ki} * U_{jk})]}$$

where

$RS_{ji}$  = the relative strength between the  $i$ th input and  $j$ th output variables

$W_{ki}$  = the weight between the  $k$ th hidden node and  $i$ th input node

$U_{jk}$  = the weight between the  $j$ th output and the  $k$ th hidden node

ABS = the sign of the absolute value

This statistic measures the strength of the relationship of the  $i$ th input and the  $j$ th output variable (expressed by numerator) to the total strength of all the input and output variables (expressed by denominator). Therefore, it can determine the relative importance between a  $i$ th input and a  $j$ th output node. In effect, this formula is similar to the form frequently used in multivariate analysis to determine the proportion of the variation of one variable in relation to all the others. This method will be applied in the empirical study in the second part of this thesis. The results in neural networks will be compared to those in MDA and Logit to measure the significance of the independent variables.

### 3.11 Incorporating the Prior Probability and Misclassification Cost

Since neural network methodologies were not originally developed for accounting classification problems, the backpropagation algorithm does not take into account the prior probabilities of each group and their misclassification cost. In the light of the bankruptcy prediction model, lack of consideration of these factors is viewed as the most severe issue compared with the selection of the appropriate classification rules and other assessment of classification accuracy [Eisenbeis, 1977]. As we have shown in Chapter Two, failure to relate the estimates of prior probabilities and misclassification costs will seriously limit the ability to make any meaningful inferences about the overall performance. Therefore, embodying these influences in the prediction model is essential.

Tam and Kiang [1992] developed a way to incorporate them into the learning algorithm by taking them into the objective function  $E$  (error function) defined as

$$E = \sum_{i=1}^2 Z_i [1/2 \sum (T_{ij} - A_{ij})^2]$$

where

$i = 1, 2$  defined as group 1 and group 2

$Z_1 = C_I p_1$

$C_I$  = the Type I error cost

$p_1$  = the prior probability of group 1

$Z_2 = C_{II} p_2$

$C_{II}$  = the Type II error cost

$p_2$  = the prior probability of group 2

For the above formula, in addition to its lack of availability in computer programmes, there exists a problem concerning incompatible results among the equal value of  $C_I \pi_1$  and  $C_{II} \pi_2$ . Specifically, if  $C_I \pi_1 = C_{II} \pi_2$ , one should expect similar results in both cases according to the above equation. But in Tam and Kiang's study, the classification accuracy is quite different. To overcome these two problems, a proposed approach will be developed in Chapter Eight. It not only considers the influences of unequal prior probability and unequal misclassification costs, but can at the same time embody the impacts of a different base rate, which is the proportion of two groups in sample in the classification model. In a bankruptcy prediction, population proportions generally bear no relation to the proportions observed in the sample. This phenomenon can cause severe bias in classification accuracy. Unfortunately, due to not considering this factor in STM as well as the ANNs, most researchers have viewed the sample proportions as estimates of prior probabilities. It is unlikely and inappropriate. The proposed approach developed in a later chapter will provide an easier solution to deal with these biases.

### 3.12 Building A Backpropagation Neural Network

The process of building a backpropagation NN consists of four stages. They are

1. Network design. 2. Network training and optimal structure selection. 3. Network validation. 4. Network prediction.



### 3.12.1 Network Design

The network design involves three steps:

- (1) Problem definition. This is the starting-point before using the neural network to solve the problem. The input variables, and the goals to be classified, recognised, predicted and generalised should be clearly identified.
- (2) Data collection. This is because neural network finds the pattern between input and output by learning the examples given. Examples or facts that have already happened should be gathered. The correct output answer in these examples or facts should also be precisely defined due to the supervised strategy of the backpropagation algorithm. For instance, in the bankruptcy prediction problem, a set of important financial ratios serve as input variables, and bankruptcy or nonbankruptcy served as the output status. Meanwhile, there must be sufficient examples of each group for the network to be able to generalise.
- (3) Network structure design. The number units in input layer, hidden layer and output layer must be specified when a neural network is built. The number of input units depends on the number of attributes involved in the decision making. Likewise, the number of output units depends on the number of classification categories in the problem. There is no specific formula to determine the optimal number of hidden units because it depends primarily upon the complexity of the problem being solved. Kolmogorov's theorem provided the suggestion about choosing the maximum number of hidden units [Caudill, 1990]. Another is the rule of thumb which recommends using the average of the number of input units and the number of output units. Finally, designing a network also involves specifying the type of activation function and other parameters.

### 3.12.2 Network Training and Optimal Structure Selection

#### Network Training

After completing the design of the network, the network will be trained. Training the network is a cyclical and interactive process. First, all the connection weights and

thresholds are randomly assigned to set the initialisation. On the other hand, a stopping rule must be defined to measure the necessary conditions for the cessation of training. During the training process, both the learning rate and the momentum term are constantly modified and vary from 0 to 1. The larger the learning rate, the more radical the change in the weight and thus the increase in the learning speed. However, the larger learning rate may lead to oscillation, and so the momentum term is added to avoid this situation. In general, a larger value of the learning rate is often selected at the beginning of the training process and gradually decreases during learning procedure. The start value of momentum has no common rule, but the lower limit is usually suggested as no less than 0.5. The training tolerance is a positive numerical value which represents the allowable variation when the actual output values are compared to the target outputs of the training facts. A lower training tolerance requires a closer match of target outputs to actual outputs. A higher training tolerance will allow more variation in the output value before errors are propagated back through the network. The usual value is set at less than 0.5. With regard to the stopping rule, two measurements are employed. One is the number of epochs which is often used as the stopping rule. The other rule is the fitting rate, which is the fraction of the input samples for which the network gives acceptable outputs. During the training, if the learning makes no progress and can not reach a satisfactory conclusion regardless the number of episodes of training, it may be necessary to go back to the network design process. The step of collecting better data or redesigning the network structure should be tried, and the training process needs to start all over again.

### **Optimal Architecture Selection**

The objective of this process is to find the simplest network which, with the least number of hidden units or hidden layers, and achieves the highest performance. Most techniques optimising the network structure involve a substantial iterative and cyclical process. In this process the tolerance rate should be specified. This parameter sets a margin around the target output values. The outputs of optimised networks within this margin are considered to be successful. Otherwise, the other architecture should be retried.

### **3.12.3 Network Validation**

After training the network, we should validate the network before using the network for future processing. Validation includes testing whether the selected optimal structure in the training process is a real optimal structure in holdout sample data, and checking whether the results of classification accuracy are satisfactory, or whether the statistics generated from the model support the assumptions related to the model. Without network validation it is difficult to justify a proposed network. With respect to the optimal structure validation, it can be solved by the techniques mentioned in the previous section. As to the latter, this involves not only concentrating on minimising the forecasting error but also checking whether any nonlinear component neglected to ensure that the best trend is fitted. In addition to the outcome of Overall error rate, Type I error rate, and Type II error rate, the residual plot can also be used for network validation. The residual plot helps to check the assumption of a random error and to detect any system pattern. Validation is essentially the same as training except that the network has the data that was not previously known. If the result of this step is good, the network is ready to use. If not, the network needs to be retrained with new information, or the network redesigned.

### **3.12.4 Network Prediction**

Network prediction is the last stage in the development of a network. After training and validation, a network can be reliably used to solve real word problems such as future prediction, pattern recognition, and so on. Unlike training and validation, the applied data have no known output, only the known input. The effectiveness of the network can not be evaluated, and nor are any corrections made at that time.

However, as time passes and more information becomes available, it is necessary to re-evaluate the performance of the network. If the performance of the network is not satisfactory, new knowledge should be built into the network, and the process of updating the network needs to be invoked until satisfactory results are obtained.

### **3.13 Summary and Conclusions**

This chapter introduced the fundamentals of an artificial neural network, the variations on the standard algorithm, the construction of the optimal architecture, and some important issues related to bankrupt prediction when using the neural network algorithm. The method of building a neural network model was also discussed in detail. The phases of building a neural network model are presented in Figure 3.13.1.

Recently, the artificial neural network has become a new challenger as a future prediction and decision-making tool. It is argued that ANNs are able to easily model any type of parametric or nonparametric process, and automatically and optimally transform the input data. Some authors advocate the artificial neural network as a replacement for statistical forecasting such as discriminant analysis or Logit regression. However, other authors are concerned that artificial neural networks might be oversold or just a fad [Chatfield, 1993]. Before making any balanced assessments of the potential of the artificial neural network, it is better to compare the advantages and disadvantages, similarities and differences between STMs and ANNs on a theoretical basis as well as in previous empirical studies. Chapter 4 will deal with the comparison between conventional statistical methods and artificial neural networks from the viewpoint of both their theoretical basis and empirical evidence.

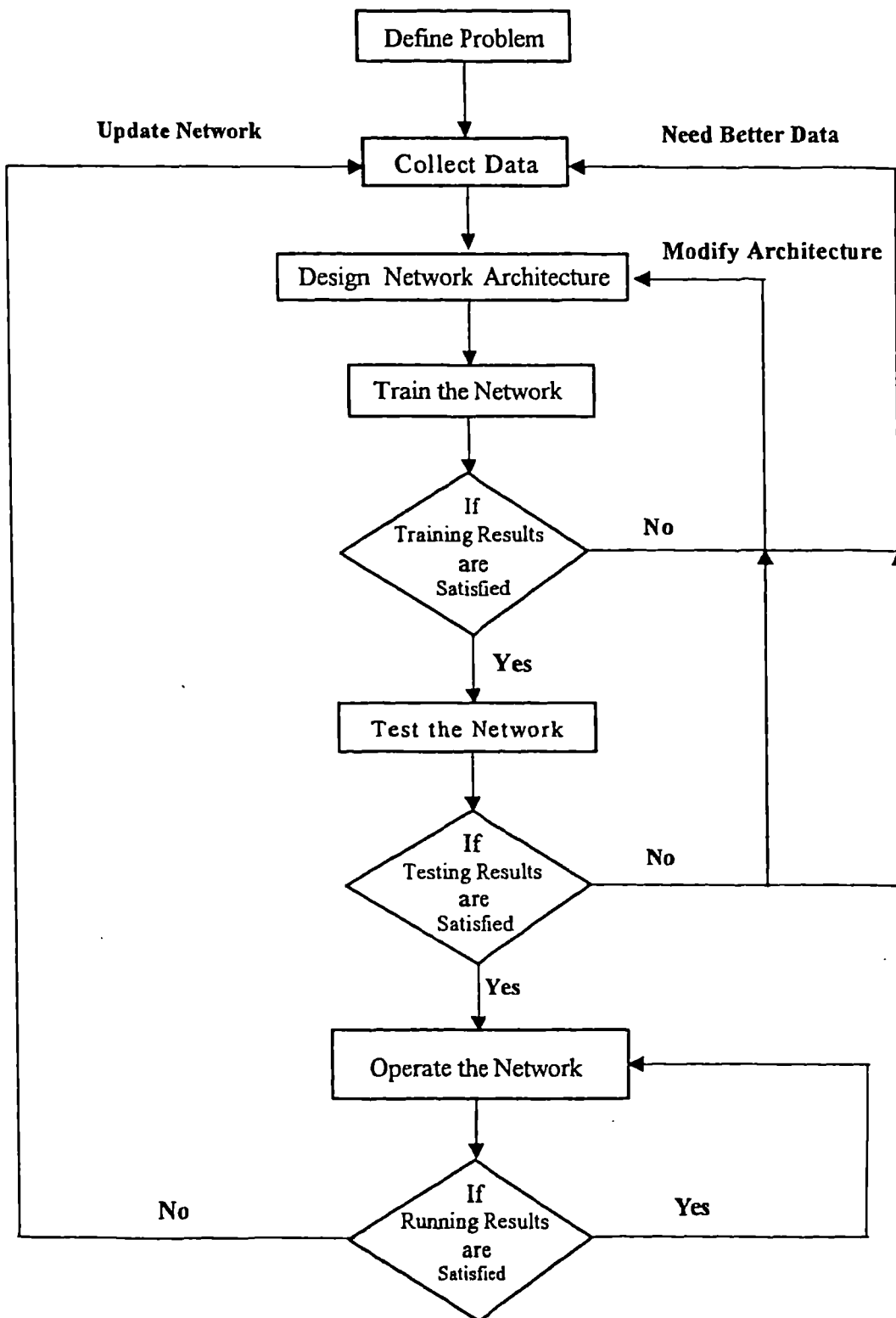


Figure 3.13.1 The Phases of Building A Neural Network

## **Chapter Four**

# **COMPARISON OF CONVENTIONAL STATISTICAL METHODS AND ARTIFICIAL NEURAL NETWORKS**

### **4.1 Introduction**

One of the neural network experts, Masters [1993], has written about ANNs and STMs as follows: "any problem that can be solved with traditional modelling or statistical methods can most likely be solved more efficiently with a neural network" (p.8). At first glance it is an exciting prospect that the neural networks can replace all existing statistical methods and outperform them. However, this comment is really a matter of dispute. In Chapter Three, the Perceptron and GDR backpropagation neural networks were introduced. But they are just two types of many neural network models. There are various artificial neural networks that are related to the nature of the task assigned to the networks. In effect the main functions of artificial neural networks are three in number

1. As models of the biological nervous system or in the analysis of intelligence in the area of neurophysiology.
2. As adaptive signal processing, image compression, speech recognition as well as automatic control application in the area of engineering or computer science.
3. As data analysis in the area of social science.

Therefore ANNs were invented as a model for human thinking, as a tool for cognitive modelling and as a data analysis technique. These cognitive models are far richer in architecture and cannot easily be compared to statistical models. The orientation should thus be to psychological, neurophysiological and even philosophical theories rather than to mathematical considerations. The way in which such networks are being trained is often part of the model and of the assumptions that are being made, and cannot therefore be replaced by a mathematically more efficient method. Thus, the comparison between ANNs and STMs should be made on the basis of ANNs as a kind of "data analysis" tool, which makes ANNs similar or identical to well-known statistical methods.

As far as data analysis is concerned, there is a basic difference between the standard "black box" philosophy of neural networks and the explicit philosophy of statistical data analysis. Statisticians depend on human intelligence, not artificial intelligence, to understand the process under study, define the problems, generate hypotheses, test assumptions, choose the best tool, diagnose problems in the model and data, and investigate the results. By contrast, neural engineers want their networks to be black boxes which do not require human intervention, but which display the inherent learning and intelligence ability between data input and prediction output. From another point of view, the conventional statistical models make rigid assumptions about model structure that they are trying to estimate from the available data. Neural networks, on the other hand, are analogous to nonparametric methods. They make no assumptions about the distribution of the data and are thus capable of letting the data speak for itself [Refenes, 1994]. However, the automation of neural networks could be sometimes an advantage as well as a disadvantage [German et al., 1992]. Within their black box, the interpretation is hindered. Ripley [1993] examined a number of case studies applying ANNs as well as STMs to a wide range of problems. He concluded that standard statistical procedures will often be at least as effective as neural networks when fair comparison is made. Sarle [1994] even went further and concluded

"It is therefore unlikely that applied statistics will be reduced to an automatic process or "expert system" in the foreseeable future. It is even more unlikely that artificial neural network will even supersede statistical methodology." (p.11).

These remarks provided by different experts seem to indicate that there exist conflicting points views on the comparison of neural networks and statistical methods even for data analysis function.

This thesis focuses mainly on financial data analysis in the problem of business failure prediction. As a consequence, comparisons of classification accuracy in bankruptcy prediction between the well-known statistical methods (MDA and Logistic regression) and the BP network models, which from some statisticians' point of view, are analogous to nonparametric, nonlinear regression models, are appropriate and necessary.

As mentioned above, for data analysis some ANN models are similar or identical to well-known statistical models, whereas the respective terminology of these two techniques shows some differences. Some definitions of these two techniques therefore need to be clarified at this point.

- (1) Independent variables in STMs are called inputs in ANNs.
- (2) Dependent variables in STMs are called desired output or target in ANNs.
- (3) Predicted values in STMs are called actual or system output in ANNs.
- (4) Residuals in STMs are called errors or biases in ANNs.
- (5) An optimisation or estimation criterion in STMs is called error function or cost function in ANNs.
- (6) Optimisation or estimation process in STMs is called training in ANNs.
- (7) Transformation in STMs is called connection or link in ANNs.
- (8) Parameter estimates in STMs model are called weights in ANNs.
- (9) In-sample performance in STMs is called convergence in ANNs.
- (10) Out-of-sample performance in STMs is called generalisation in ANNs.

## **4.2. Theoretical Comparisons of ANNs and STMs**

Five aspects of comparisons will be made between neural networks and conventional statistical models. They are (1) model formulation and problem solving procedure; (2) nonlinearity vs. linearity boundary building; (3) statistical testing and interpreting; (4) model generalisation, and (5) adaptability. Through these comparisons the limitations and benefits of each of them can be further understood.

### **4.2.1 Model Formulation and Problem Solving Procedure**

#### **4.2.1.1 Perceptron vs. Statistical Models**

A mathematical expression for a linear model is as follows



$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_n X_{in} + \varepsilon_i = \sum_{j=0}^n \beta_j X_{ij} + \varepsilon_i$$

where

$Y_i$  : the value of dependent variable for the  $i$ th observation

$X_{ij}$  : the value of the  $j$ th independent variable for the  $i$ th observation

$\beta_j$  : the unknown parameter for the  $j$ th independent variable

$\varepsilon_i$  : the residual term which is assumed to be  $E(e)=0$ ,  $V(e)$  is constant

Each linear model has a different objective in analysing the data. For the purpose of discriminant analysis the task is to check whether the mean vectors from two different groups have the same values, and to classify a new object into one of groups correctly. On the other hand, the aim of regression analysis is to estimate the coefficient of the model and to find out the functional relationship between the independent variables and dependent variables. In MDA the assumptions involved in the independent variables are multivariate normal distribution and equal variance-covariance matrices across groups. In conditional probability regression models, a distribution assumption for the residual term such as logistic (Logit) or normal (Probit) distribution is required.

Perceptron has only an input layer and an output layer. Assume that the transfer function of the output node is linear, Perceptron has its output as in the following equation

$$Y = f(\sum W_i X_i)$$

Since  $f$  is assumed to be a linear function, it has the same form as the linear regression function [Weisberg, 1985]; [Myers, 1986], possibly multiple or multivariate, as shown in Figure 4.2.1.

$$Y = \sum bX$$

Although they have almost the same meaning in terms of a mathematical equation, one major difference between the two models is the estimation process. In the network's learning process, when the weights are estimated, the data is assumed to come into the system sequentially. The weight is then adjusted and estimated according to a certain rule. However, in the regression model it is assumed that a fixed set of data is given. Therefore, the final estimate of the neural network may be the locally minimising mean square error function instead of the globally minimising mean square error function, due to its features in the training process.

When the Perceptron has a nonlinear transfer function such as sigmoid, it is a logistic regression model (Logit) [Hosmer and Lemeshow, 1989]. The relationship between the input and output in this case becomes

$$Y = \frac{1}{1 + \exp(-\sum W_i X_i)} \text{ or } \sum W_i X_i = \ln\left(\frac{Y}{1-Y}\right)$$

It has the same form as the logistic equation. The only difference between them is that the output of dependent variable  $Y$  in the Logit usually represents the probability of occurring rather than a predicted value in the ANN. Figure 4.2.2 illustrates this situation. A Perceptron with a threshold transfer function is a linear discriminant function [Hand, 1981] [McLachlan, 1992]; [Weiss and Kulikowski, 1991]. The problem addressed by linear discriminant analysis is to find a set of weights so that the variability within each group is as small as possible, whereas the variability between different groups is as large as possible. According to the theory developed by Fisher [1936] and later extended by others, the weights to be used for calculating the Z score should involve two matrices A and B. Matrix A is set for the collection of all the objects belonging to all groups, which indicates the "total" sum of squares, while matrix B is set up by finding the sum of squares and cross-products for each group separately and then adding up these matrices, which indicates the sum of squares "within" the group. The difference of A and B is denoted by C. Based on the idea behind the theoretical solution of the LDF, it turns out that the eigenvectors of the matrix given by  $B \times C$  would be the weights that would cause the smallest percentage of cases to be assigned to wrong groups. If the number of groups is  $m$ , and the number of variables is  $n$ , the number of such discriminant functions that can be calculated is the smaller of  $m-1$  and  $n$ . With two groups, only one discriminant function would be used, with a cutoff point so that observations with scores above the cutoff score would be assigned to one group, and otherwise would be assigned to the other group. These features are closely related to those that are used by a Perceptron with a linear function. The nature of threshold in the transfer function in a neural network is equivalent to the cutoff point in LDF. Accordingly, with the same concept of the multiple outcomes in the output, the Perceptron becomes a multiple discriminant function. This relationship is shown in Figure 4.2.3.

Despite the similarities mentioned above, there are some differences between Perceptron and LDF. In contrast to a linear discriminant function, the Perceptron network is inherently nonparametric. No assumption of normality are made with respect to the population, nor are means or covariance necessary. However, compared to the estimation process, Perceptron networks may lack stability, especially in those instances where there are a small number of training examples [Weiss and Kulikowski, 1991]. In such cases, there are too few cases to specify the distribution completely, allowing for the instability.

#### 4.2.1.2 Multilayer Network vs. Statistical Models

In addition to an input and output layer, an extra hidden layer introduced to the network produces a multilayer network (MLP). There is no need for a linear transfer function in each hidden node, since the relationship between the input and output in this type of neural network can be simply expressed by Perceptron with the linear transfer function.

If the model includes estimated weights between the inputs and the hidden layer, and the hidden layer uses nonlinear transfer functions, the model becomes genuinely nonlinear, which means that the estimated parameters have nonlinear form. The one hidden layer network with sigmoidal transfer function is a logistic nonlinear regression. To explain this easily, assume a two-input network with only one node in the hidden layer as shown in Figure 4.2.4.

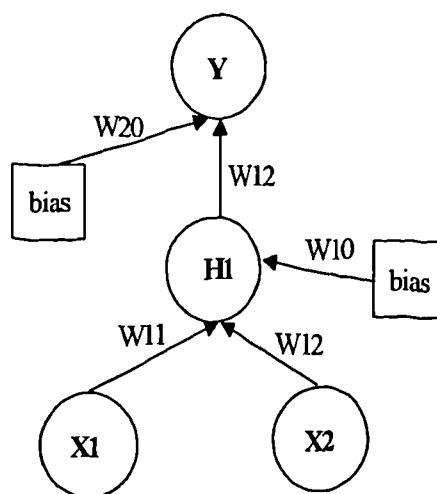


Figure 4.2.4 A Network with One Hidden Node

The general expression for the output of one hidden layer is

$$Y = f_k(\sum_j W_{kj} f_j(W_{ji} X_i)) \quad i = \text{an input node, } j = \text{a hidden node, } k = \text{a output node}$$

If the transfer function is sigmoid function, then the output of the network is

$$Y = \frac{1}{1 + \exp[-(W_{20} + W_{21} H_1)]} \quad (4.2.1)$$

And the output from the hidden node is

$$H_1 = \frac{1}{W_{21}} \left( \ln\left(\frac{Y}{1-Y}\right) - W_{20} \right) \quad (4.2.2)$$

The value of the hidden node can also be expressed

$$H_1 = \frac{1}{1 + \exp[-(W_{10} + W_{11} X_1 + W_{12} X_2)]} \quad (4.2.3)$$

If we combine equation (4.2.2) and (4.2.3), then

$$\frac{1}{W_{21}} \left( \ln\left(\frac{Y}{1-Y}\right) - W_{20} \right) = \frac{1}{1 + \exp[-(W_{10} + W_{11} X_1 + W_{12} X_2)]} \quad (4.2.4)$$

The relationship between Y and  $\ln(Y/(1-Y))$  is very close to linear over much of the range 0 to 1 as the Figure 4.2.5

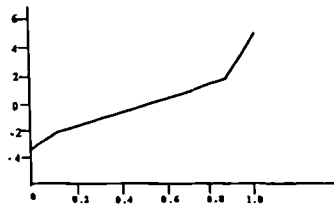


Figure 4.2.5 A Graph of Y vs.  $\ln(Y/(1-Y))$

Due to the complexity of the relationship, it is difficult to find the effect of  $X_i$  on Y directly. Suppose Y is the linear approximation of  $\ln(Y/(1-Y))$ , then the equation (4.2.4) can be rewritten

$$Y^* = W_{20} + \frac{W_{21}}{1 + \exp[-(W_{10} + W_{11} X_1 + W_{12} X_2)]}$$

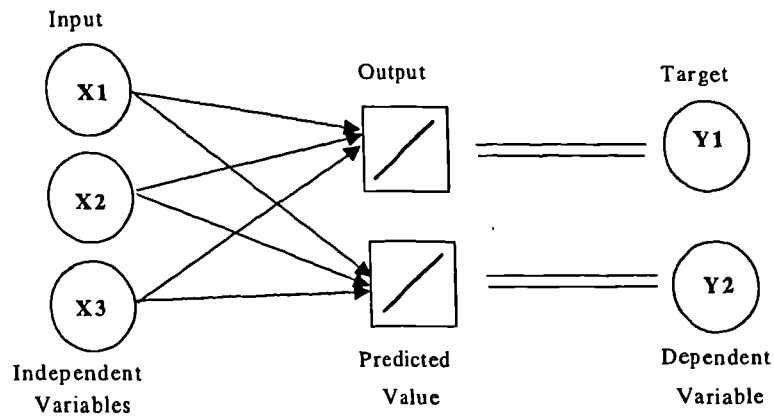


Figure 4.2.1 Simple Linear Perceptron = Multiple Linear Regression

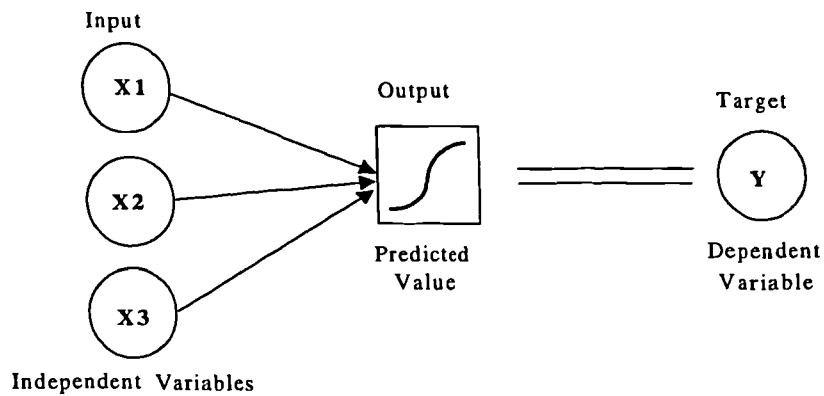


Figure 4.2.2 Simple NonLinear Perceptron = Logistic Regression

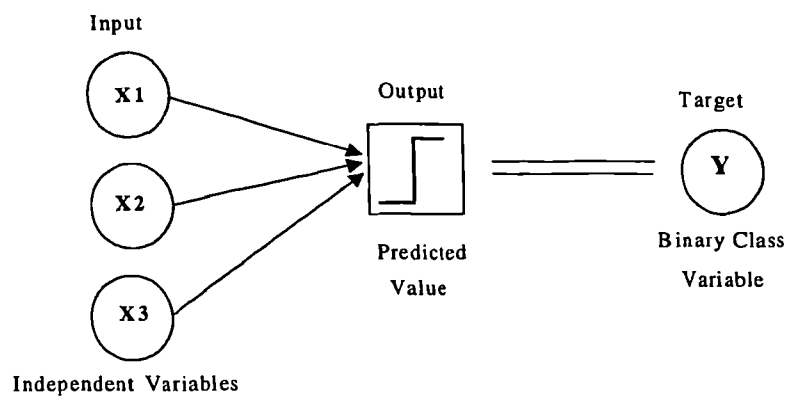
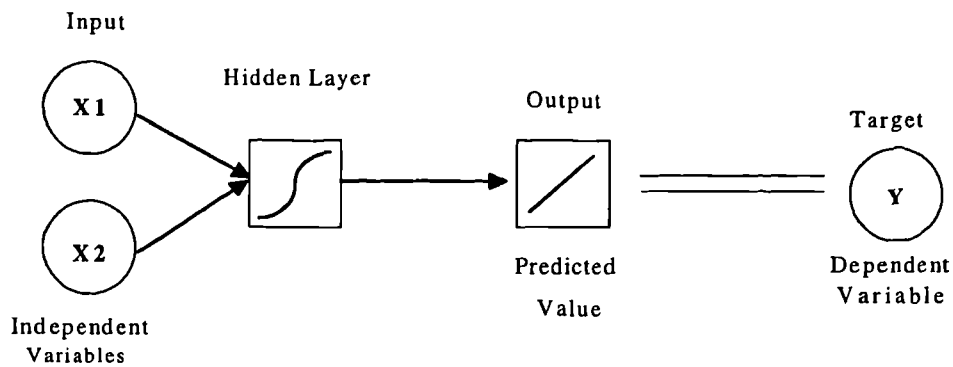
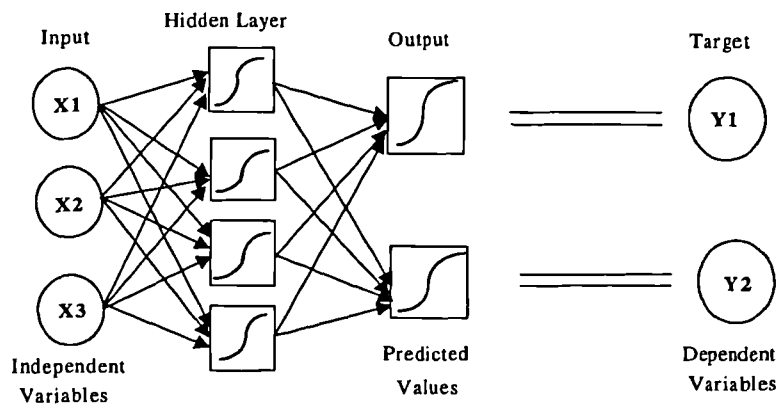


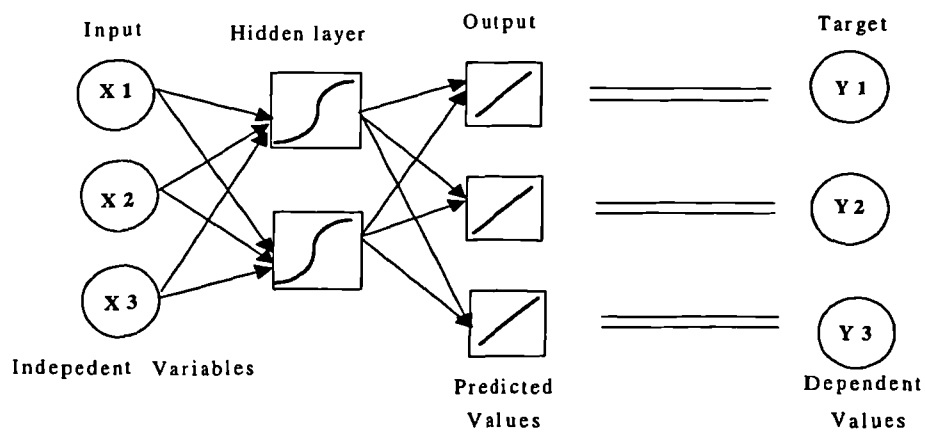
Figure 4.2.3 Perceptron with Threshold = Linear Discriminant Function



**Figure 4.2.6 Multilayer Perceptron = Simple Nonlinear Regression**



**Figure 4.2.7 Multilayer Perceptron = Multivariate Multiple Nonlinear Regression**



**Figure 4.2.8 Multilayer Perceptron = Nonlinear Regression Again**

As a result the functional relationship of  $X_1$  on  $Y^*$  given  $X_2$  follows the logistic relationship. The resulting model is displayed as Figure 4.2.6.

An multilayer network with multiple inputs and outputs becomes multiple nonlinear regression. Figure 4.2.7 and 4.2.8 indicate this situations.

In effect, if we further explore the nature of MLP with the BP algorithm, backpropagation is an iterative gradient technique that is similar in many ways to the Newton-Raphson technique used in the maximum likelihood estimation of the Logit model. In this respect the neural network with sigmoidal function can be regarded as a weighted multi-logistic regression model. The basic premise underlying the backpropagation algorithm is that each of the network connection weights is, to some degree, responsible for the final output error.

### **Comments on the Model Formulation and Problem Solving between ANNs and STMs**

A linear model such as linear discriminant function or logistic regression is deriving a function relating explanatory variables  $X$  vector to an output  $Y$ . The problem is solved by minimising some error measure or maximising some likelihood function, and the fitted coefficients attached to each explanatory variable are thus obtained; whereas the neural network solves the problem in terms of a heuristic search [Triguerios and Taffler, 1995].

The neural network starts the search for minimum error by setting the weights (equivalent to the coefficients in statistical model) randomly. An observation in the data set is then randomly chosen to present through the input layer to the output layer. The weights are then updated to reduce the error according to the difference between the desired output and the actual output. After all observations have presented in the network, the updated weights tend asymptotically towards describing the underlying relationship. The iterative procedure will find a minimum in the cost function and will thus yield results similar to those generated by conventional statistical methods. However, there are some distinctions

- (1) Conventional statistical techniques require some assumptions in deriving the model.

MDA needs normality in the distributions of independent variables and the homogeneity of variance-covariance across groups. Logit (logistic regression) is also restricted by the logistic distribution assumption on error terms. However, the

neural network algorithm does not rely on any specific distribution of the variables and the assumptions of statistical models.

- (2) Instead of fitting the desired relationship by means of just one unique or logistic function, the neural network may fit several linear functions in the initial step through the hidden layer, which are then fed into the other function in the other hidden layer, in order to provide the overall output value. This procedure produces substantial differences in explaining the interactions effects, and thus in the quality of classification, between neural network methodology and traditional statistical approaches.
- (3) Since neural networks do not share the assumption of normal distribution in independent variables, and do not impose a linearity constraint in MDA, nonnumerical data which denotes, for example, firm size, industry, nationality or spurious market behaviour can be used as numeric input.
- (4) The neural network methodology tends to be good at capturing multivariate data that distinguish various outcomes, while the conventional statistical models such as MDA and Logit procedure focus on capturing a single pattern, and break down the explanatory variable estimate into parts which can be separately forecast, rather than on identifying discriminating patterns. In another words, the prediction ability in neural networks contrasts with the decomposition capability in conventional statistical methods [Gorr, 1994].

#### **4.2.2 Nonlinearity vs. Linear Boundary Building**

The second comparison between ANNs and STMs is in terms of the boundaries they form. The main advantage of a multi-layered neural network probably lies in its ability to form the complex boundary underlying the nonlinear attributes in the problem. In Perceptron networks with a threshold transfer function, which is another form of linear discriminant analysis, input nodes connect directly to output and each connection has a weight attached to it. It should be noted that the Perceptron with no hidden layer can only cope with the linear problem or some specific function through the defined transfer function. Even the Perceptron model with a nonlinear transfer function is not genuinely nonlinear. It can only

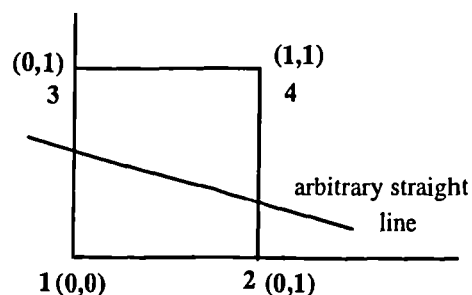


solve the problem if it has the same attributes with the transfer function defined. This is the reason why the Perceptron cannot solve the nonlinear XOR problem. Linear discriminant analysis encounters the same difficulties with Perceptron since they are similar concepts in terms of certain design elements. MDA requires that the decision set used to distinguish between groups must be linearly separable. For a single input it is possible to choose a cutoff point that is above or below this value, and thus an observation can be classified into the correct group. However, when two or more input variables are considered together, they can not be completely distinguished just by a line or by a plane because multivariate inputs form a nonlinear region in most cases. This makes MDA incompatible with a complex decision boundary. On the other hand, the multilayer Perceptron model is a general purpose, flexible nonlinear model. It is a universal approximation approach, and can fit any function to any degree of accuracy if appropriate hidden nodes are chosen even with little knowledge about the form of the relationship between the independent and dependent variables [White, 1992].

One widely known limitation of the Perceptron is that it cannot compute the XOR function, which is of paramount importance in pattern recognition applications. Its crucial shortcoming is shown below. Consider the following problem

$X_1$	$X_2$	Y	Group
0	0	0	1
0	1	1	2
1	0	1	2
1	1	0	1

This problem is equivalent to the four corners of a square



The first and fourth points are in group 1, the second and the third points are in group 2. A simple Perceptron with only input and output layers computes a linear sum of these input and output values 0, 1 as a output. In geometry this means drawing a straight line that tries to partition this square into two regions. The position and direction of the straight line are determined by the weights on the input connections. But whatever we draw, we can not divided the square into two regions so that (0, 1) and (1, 0) end up in one region and (0, 0), (1, 1) end up in the other. This phenomenon can also be exhibited by a linear function as in Fisher's linear discriminant function. Consider a linear function in the following way

$$Z = a_0 + a_1X_1 + a_2X_2$$

Set a cutoff point  $Z^*$  which is used as classifying group 1 or group 2. The criterion is when  $Z < Z^*$  (cutoff point), the observation (point) is classified in group 1; otherwise in group 2.

(1) If point 1,  $(X_1, X_2) = (0, 0)$  is claimed it should belong to group 1, that means  $a_0 = Z < Z^*$

(2) If point 2,  $(X_1, X_2) = (0, 1)$ , claiming it should belong to group 2, it is  $a_0 + a_2 = Z > Z^*$ .

What implied in (1) and (2) is  $a_2 > 0$ .

(3) If point 3,  $(X_1, X_2) = (1, 0)$  also be claimed belonging to group 2, it leads to  $a_1 > 0$  since  $a_0 + a_1 = Z > Z^*$

while

(4) When point 4,  $(X_1, X_2) = (1, 1)$  derives  $y = a_0 + a_1 + a_2$ . From (1) (2) and (3), the value of  $Z$  must be greater than  $Z^*$  because  $a_1, a_2$  which we just derived are all greater than 0.

Hence, it never cannot be classified as group 1.

The dilemma occurring in Perceptron or LDF can be easily solved by introducing a hidden layer between the input and output layers. Since a hidden unit in the hidden layer allows the network to partition the input space into arbitrary nonlinear regions, it provides powerful computational effectiveness for this complex problem.

In the XOR case, larger ANN model which contains one hidden layer with the appropriate weights and thresholds is represented in Figure 4.2.9. In this two hidden nodes network, each hidden node can partition the input space in a different way. The output then computes a linear combination of these partitionings to solve the problem as displayed

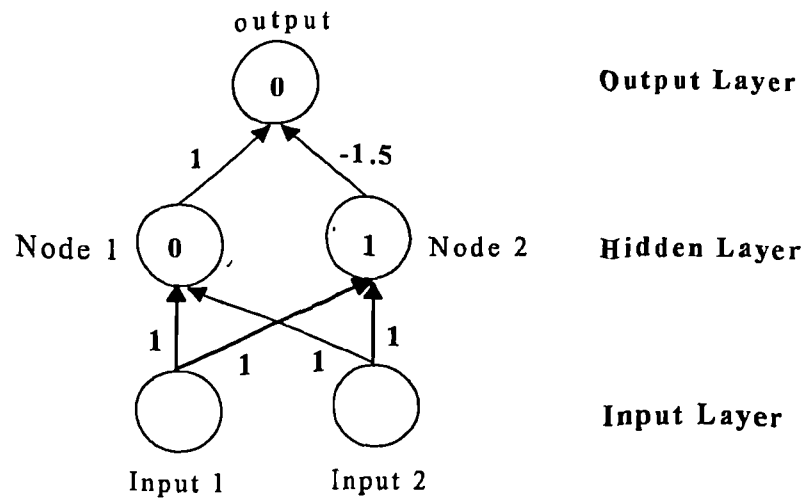


Figure 4.2.9 Multilayer Neural Network to Solve XOR Problem

Using notations:

$W_{11}^{ih}$  : weight connecting input variable 1 (INP1) with hidden node 1

$W_{12}^{ih}$  : weight connecting input variable 1 (INP1) with hidden node 2

$W_{21}^{ih}$  : weight connecting input variable 2 (INP2) with hidden node 1

$W_{22}^{ih}$  : weight connecting input variable 2 (INP2) with hidden node 2

$W_{1o}^{ho}$  : weight connecting hidden node 1 with output node

$W_{2o}^{ho}$  : weight connecting hidden node 2 with output node

$WS_1$ : weighted sum in hidden node 1= $INP1 \times W_{11}^{ih} + INP2 \times W_{21}^{ih}$

$WS_2$ : weighted sum in hidden node 2= $INP1 \times W_{12}^{ih} + INP2 \times W_{22}^{ih}$

$H_1$ : the output in hidden node 1 (active or not)

$H_2$ : the output in hidden node 2 (active or not)

$V_o$ : the value in output layer= $H_1 \times W_{1o}^{ho} + H_2 \times W_{2o}^{ho}$

The number in each node is its threshold

We calculate the actual output value and the desired output values at each point

Point	$WS_1$	$H_1$	$WS_2$	$H_2$	$V_o$	The actual output value	The desired output value
(0, 0)	0	0(inactive)	0	0 (inactive)	0	0	0
(0, 1)	1	1(active)	1	0 (inactive)	1	1	1
(1, 0)	1	1(active)	1	0 (inactive)	1	1	1
(1, 1)	2	1(active)	2	1 (active)	-0.5	0	0

This network correctly solves the XOR problem. When neither input node  $(0, 0)$  is active and neither hidden node is active, the output node also is off. When either a single input node is on  $(0, 1)$  or  $(1, 0)$ , the hidden node is on and the actual output is on. If both input nodes are on, the hidden nodes are activated. However, the large negative weight from hidden node 2 to the output unit is greater than the weight from hidden node 1 to the output, and the output node turns off. From another perspective, when the interaction effect is larger than the main effect, it will be classified in group 2, otherwise in group 1.

By adopting a geometric viewpoint, as shown in Figure 4.2.10, the first hidden node partitions the space so that it is activated when either input node  $(0, 1)$ ,  $(1, 0)$  or both  $(1, 1)$  are active; whereas the second hidden node becomes active only when both input nodes  $(1, 1)$  are active. It has a stronger inhibitory (negative) influence on the output node than the excitatory influence of the first hidden node.

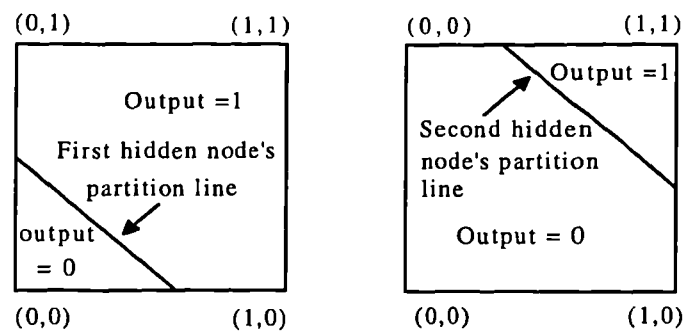


Figure 4.2.10 The Decision Line in ANN to Solve XOR Problem

The hidden node acts as a feature detector, filtering information, and determines whether it should be used or not. In this way the multilayer ANN can divide the linearly inseparable data and classify a group correctly. The results above seem to demonstrate that the Fisher linear discriminant function not only suffers from its restrictive assumptions of multivariate normality of independent variables and equal variance-covariance matrices in the two different groups, but also creates methodological problems due to the linear monotonicity [Yoon et al., 1993], so-called reversal of the likelihood function [Moore, 1973]. Thus, the LDF may always misclassify some certain types of problem like XOR. As a consequence, if a problem contains more complex interactions between independent variables, the MLP neural network methodology may provide a more powerful explanation in classification accuracy than the linear model because of its accurate frontier building.

### **4.2.3 Statistical Testing and Interpretation**

#### **4.2.3.1 Model Fitting**

The third comparison between ANNs and STMs is related to statistical testing and interpretation.

The overall significance testing of the model, and the relative importance evaluation of independent variables are essential parts of statistical methods. In MDA the F statistic is used to examine the contribution of classification variables in the model, namely, by assessing the appropriateness of model fitting. In other words, if the null hypothesis that none of the variables improves the classification based on chance is true, then the function measurement follows a F distribution. Likewise, the test statistic used in Logit procedure to assess overall fit is the -2 times Log likelihood ratio. This measure similarly tests the null hypothesis that the independent variables have no impact on the prediction of classification. If this hypothesis is true, then the test static follows a chi-square distribution. On the other hand such significance tests are developed with difficulty in a neural network because of its distribution free optimisation attribute. Gorr et al. [1994] criticised this issue in the neural network algorithm

"... since the ANN [artificial neural network] model form is non-linear in the model coefficients, the normal probability model is not applicable. Consequently, ANNs do not have parametric statistical properties (e.g. they do not have individual coefficient of model significance tests based on the t or F distribution." (p.19).

The issue of the lack of such significance assessment could mean that some of the benefits of the neural network approach will become less evident. [Trigueros and Taffler, 1995].

#### **4.2.3.2 Model Interpretation**

Several approaches that attempt to determine the relative importance of individual variables have been proposed and used in multivariate discriminant analysis (MDA). They include (1) the standardised coefficient method, which involves the multiplication of pooled standard deviation and its corresponding coefficient. The larger the standardised discriminant

coefficient value, the greater its contribution; (2) The Mosteller and Wallace method [1963], which measures the importance of a particular variable in terms of the proportion of Mahanobis's distance  $D^2$ ; (3) The Conditional deletion method which involves removing one variable from the entire  $k$  variable set at one time. The residual Wilks' Lambda of the variables is ordered according to the resulting reduction in overall discriminating power measured by  $F$  statistic in the rest of  $k-1$  variables. The highest residual Wilks' Lambda is considered as the most significant in the  $k$  variable discriminant function.

In the Logit procedure, tests of significance of individual model coefficients are commonly used either as the  $Z$  statistics or as Wald chi-square statistics. Regardless of the methods employed in conventional statistical techniques, they all tend to be logical and offer a direct interpretation in explaining the individual contribution. As we have shown in Chapter Three, researchers have developed some approaches concerning the estimation of significance of input variables in a neural network. However, owing to the input unit being indirectly connected to an output unit through hidden nodes, the outputs can hardly be directly interpretable. Gorr [1994] commented that "...insight from the behaviour of individual model components explaining estimates or forecasts are difficult to obtain" (p.2). Because the analysis of weights gained from the final ANNs output is complex and difficult to interpret, the neural network seems to be unable to assist in clearly understanding the underlying behaviour between input and output. This makes it impossible to identify the causes of the errors or defective responses. However, Chua [1986] indicated that in scientific method, the objective of the research process in the financial or accounting area should be not only the apparent performance of a model but also a better understanding of the underlying accounting issues of concern to the researchers. Triguerios and Taffler [1995] also supported this point and emphasised that "the analytical tool used is not of intrinsic interest itself but only a means for elucidating the underlying phenomena" (p.11). Altman et al. [1994], after carefully undertaking a comparative study on corporate distress diagnosis between discriminant analysis and neural network methodology, concluded that conventional discriminant analysis proves to be a very effective tool. It has the advantage, especially for the financial analysis and economic evaluation, of making the underlying model transparent and easy to interpret.

#### 4.2.4 Model Generalisation

Generalisation is our fourth concern of comparison between ANNs and STMs.

The multilayer neural network with enough hidden units can approximate any relationship between input and output variables. This attribute, although providing the advantage of formulation of a nonlinear classification boundary, also leads to a tendency towards overfitting. Since a network concentrates on the best fit of the function relationship in the training processing, it may lose the generalisation out of sample testing data. There seems to be a trade-off between memorisation and generalisation in a neural network when it produces "optimal" network designs in some problems.

In particular, when a network is applied in the accounting area, which is viewed as a quite distinct cognitive domain from those for which an ANN technique was originally developed, the MLP network may be too powerful an instrument for the relatively simple relationships conventionally founding in accounting [Trigueiros and Taffler, 1995]. Altman et al. [1994] criticised the long processing time and the arduous trial and error process required to discover the best model structure and simultaneously to avoid the trap of over-fitting. Even though a huge effort has already been made, the out-of-sample accuracy is still usually the victim of this overfitting problem. Trigueiros and Taffler [1995] also pointed out that "change in the output variable are not monotonically related to small perturbations in input variables considered one at the time. This phenomenon is consistent with the existence of a degree of overfitting and sample bias in the derived model" (p.10). Moreover, from the viewpoint of certain researchers, the better learning performance of ANNs in comparison to STMs is only due to their suspicious overfitting feature. Overfitting data has been the downfall of complex model forms used in forecasting, resulting in poor generalisation accuracy [Gorr, 1994]. Gorr et al. [1994] also stated

"Where relatively few explanatory measures are available for making predictions, simple models are often the best, and perhaps no amount of sophisticated methodology will make any improvement. ... Thus, in cases where there is no underlying structure in the available data, ANN is simply not going to perform any better than the simpler model." (p.19).

#### **4.2.5 Adaptability**

The last comparison between ANNs and STMs is in terms of adaptability.

A model is only useful for predictive purposes if the underlying relationships and parameters are stable over time. Otherwise it will only be valid for the sample period and it cannot be extrapolated into a subsequent period with the same expected performance [Altman and Eisenbeis, 1978]. One of the attractive properties of neural networks is that they allow adaptive adjustment to the predictive model as new observations enter. When the underlying distributions are changing, past information is not totally ignored but is gradually reduced in importance as new examples are fed into the network. This adaptive learning process is compatible with phenomena in the real world and is a very important aspect of ANN's effectiveness. However, statistical methods need batch update, and the entire training set is used to construct a new model when new examples are applied. They cannot adjust themselves to the feature of a gradual transformation of the environment since statistical models assume that old and new examples are equally valid.

### **4.3 Empirical Comparisons of ANNs and STMs**

The comparative analysis between commonly used STMs and newly developed ANNs for classification accuracy has only been conducted in the last few years. We summarised the representative work in Table 4.3.1.

As can be seen in the table, most of these studies are biased. For instance, the process of selection of financial ratios is a probably a major problem. In many papers financial ratios used according to previous studies can represent significant predictor variables neither for statistical approaches nor for the network models. For some others, the use of stepwise procedure favours the MDA, since it formulates an optimal discriminant function that also maximises the 'distance' between the group. In Coats and Fant's work, we could criticise their choosing the auditors' qualification as classification accuracy standards, because the auditing opinion is an inexact and subjective process, and thus their standard is possibly an incorrect indicator of distress. In addition, the often adoption of small sample sizes also cause our doubt about the reliability of their results. Moreover, the prediction risk (true



Table 4.3.1 Summary of Previous Comparative Studies of ANNs and STMs

Study	Compared Techniques	Outperformed	Some Natures of Research Design or the Significant Results
Gallinari et al. [1988]	BP ANN vs. MDA	ANN is the same as MDA ANN	Using a linear transfer function Using a nonlinear transfer function
Dutta and Shekhar [1988]	BP ANN vs. Logit	ANN	The results showed that the number of hidden layers and units is not a crucial element in ANN design
Ahlt et al. [1989]	BP ANN, Mmaximum-Likelihood Classifier and K-Nearest-Neighbor	Maximum-Likelihood Classifier and Nearest-Neighbor Technique	ANN is insensitive to the noise level
Lippmann and Beckman [1989]	BP ANN (to approximate various nonlinear functions)	ANN can successfully catch the nonlinear function	Using an ANN with two hidden layers with 20 nodes in the first layer and 5 nodes in the second hidden layer
Odom and Sharda [1989]	BP ANN vs. MDA	No conclusive evidence	Bankrupt firms are more accurately classified in ANN
Shadmehr and D'Argenio [1990]	BP ANN vs. Maximum Likelihood Estimator and Bayesian Estimator	ANN	The neural network can be an alternative to Bayesians posterior probability when the prior probability is unknown.
Denton et al. [1990]	BP ANN, MDA, QDA and Linear Programming	ANN	The ANN is quite robust in the presence of an outlier
Bell, Ribar and Verchio [1990]	BP ANN vs. Logit	No conclusive evidence	Using a trial and error process to determine the best ANN topology ANN
Erdeben and Koch [1991] [1992]	BP ANN vs. MDA	MDA (1991) No significant difference (1992)	72 financial ratios were reduced to the 4 most relevant ratios by factor and stepwise discriminant procedure
Utun and Moody [1991]	BP ANN vs. Linear Regression	ANN	The elimination of input variables and the choice of best network was carefully explored

Table 4.3.1 Summary of Previous Comparative Studies of ANNs and STMs  
(continue)

Study	Compared Techniques	Outperformed	Some Natures of Research Design or the Significant Results
Salchenberger, Cinar and Lash [1992]	ANN vs. Logit	ANN	The number of hidden unit is based on the rule of thumb, which suggests that the number of nodes in the hidden layer should be 75% of the number of nodes in the input layer
Tam and Kiang [1992]	BP ANN, MDA, Logit and ID3	ANN (but has shown overfitting problem)	The topology of ANN was based on the trial and error procedure
Coats and Fant [1993]	BP ANN vs. MDA	No significant difference in Type II error. But ANN is superior to MDA in Type I error	Cascor algorithm was utilised to self-determine the number of hidden nodes
Kim, Weistroffer and Redmond [1993]	BP ANN, QDA, Logit, Regression and Rule-Based System	ANN	Even the degree of inaccuracy in the ANN is better than other methods whenever an incorrect is given
Yoon, Swales and Margavio [1993]	BP ANN vs. QDA	ANN	The results showed that the more hidden layers there were in the model, the higher was classification ability
Altman, Marco and Varetto [1994]	BP ANN vs. MDA	No conclusive evidence	This study is more carefully designed and explored the nature of ANN more extensively than any other similar research
Wilson and Sharda [1994]	BP ANN vs. MDA	ANN	Using multiple subsamples Using different base rate combinations between training and testing data sets
Refenes, Zapranis, and Francis [1994]	BP ANN vs. Regression	ANN	The significance of inputs was presented as an important component of the analysis

generalisation ability) and architecture selection for neural networks have not been fully investigated in almost all research presented above.

However, there are two of them worth mentioning, one is the work of Altman et al. [1994]. Altman, a pioneer using MDA technique, was also attracted by the appeal of the neural network, and made a comparison between the LDF and BP algorithm for Italian corporate distress classification and prediction [Altman et al. 1994]. This study was more carefully designed and explored the nature of neural network more extensively than any other similar research. Four subjects were highlighted in the study: (1) First, they found out the capability of a neural network to reproduce the accuracy of numeric values of the scores obtained using linear discriminant analysis, using different input ratios from those employed in discriminant analysis. The experiment was performed using networks varying the number of input ratios, and the number of hidden layers and the number of hidden nodes in order to verify the networks' capacity to approximate the discriminant analysis linear function. This involves checking the neural network's ability for adaptation and simplification by examining if these approximations can be obtained with a smaller set of inputs or less effort than these was for the estimation of the discriminant function. The best result was obtained for a four-layer network with 10-10-4-1 configuration. 808 companies, 404 each from healthy and unsound groups, were trained and interrupted after 1000 learning cycles to adjust the weights. The results showed that the network's capacity for adaptation was encouraging but at the expense of complexity in architecture and more machine-hours. Moreover, the input indicators built in the network to replicate the discriminant function are completely different from those included in the functions. (2) Second, they discovered the capability of the ANN to separate the samples between bankrupt and healthy companies. Networks with varying degrees of complexity were trained. The most satisfactory results were obtained with a four-layer network with 15-15-6-2 configuration. This configuration provided the classification ability of 97.7% of healthy and 97% of bankrupt companies. This result, although outperforming the recognition rates obtained by the MDA, used a higher number of input indicators: fifteen as opposed to nine, and with erratic learning behaviour. More importantly, the ANN indicated a lower ability for generalisation than with the traditional discriminant function. This conclusion was reinforced by the results obtained on the other independent samples of 302 companies. (3) Third, they investigated the ANN's capacity to respond to the change in company performance over time. This experiment made use of the

logic of networks with memories on the input. Inputs including the entire three-year historical series of indicators were used, and the network was trained to consider all the data available about the company at the same time, just like a financial analyst examining the historical time series of financial statements. The overall accuracy of networks with memories is over 99% both for healthy and unsound companies. In the light of the results obtained, the ANN seemed to have a great potential for making predictive ability sensitive to the passing of time in identifying distressed companies. (4) The last subject to be investigated was to check the capacity of networks to separate the three categories of company: healthy, vulnerable and unsound. The results of the two-output unit networks trained simultaneously to recognise the three groups of business financial status are promising. However, taking into account the results obtained in the control periods and in the holdout samples, MDA was deemed to be better.

The conclusions reached by this study are interesting and may be summarised

- (1) On the whole, the MDA is not worse when compared to ANNs. The linear form, albeit with the limitations of its ability to perform well, ensures consistent behaviour for any type of variable and can interpret *the model's operating logic on the basis of* the coefficients.
- (2) Complex networks with large hidden nodes tend to have a better classification with heterogeneous observations, but suffer from long training time and the adoption of oscillating or nonconvergent behaviour as well as the sacrifice of the overfitting trap.
- (3) Integration of neural network and discriminant functions could offer the better future research direction. In relation to a problem which is less clear and more complicated, it may be helpful that the neural network is employed to extract the families as a simple structure before putting them into discriminant analysis.

The other important work is Refenes's study. In contrast to Yoon's study, Refenes et al. [1994] examined the performance of neural networks and regression models for forecasting within APT (arbitrage pricing theory) model for stock ranking. As a matter of fact, Refenes tried to construct the portfolio through rating the stocks. Hence the framework provided by him was more complex and dynamic than that of Yoon et al. In this experiment 143 stocks were chosen. The data set period was from May 1985 to December 1991 on a monthly

basis. Three undefined factors A, B, C were extracted from the balance sheet of the companies in the universe of the UK stocks to be used as explanatory variables. The output Y represented the outperformance of each stock instead of the binary classification outcome. The learning algorithm employed in the neural network is simple GDR backpropagation just as in other binary or multiple classification applications. Network architecture was set up with a 3-32-16-1 structure. This is a two hidden layer configuration with an unusually large than the number of input variables. This choice was a product of trial and error. The result showed that even a simple designed network far outperformed regression analysis for stock ranking. They claimed that the smooth interpolation properties allowed the ANN to fit the model much better and to generalise more successfully for this complicated financial problem. The comparison in this paper was made on the basis of convergence, generalisation stability and sensitivity analyses—more extensive and more reliable. We think this work is a good example of solving ranking not just classification.

Finally, I would like to close this section with an interesting comment by Aharonian [1992]

"There is a paper by White [1988] at UCLA who tried using neural networks to forecast the closing price of IBM's stock. His paper, and a few others I have, all tend to conclude that using neural networks for such activities is no more accurate than using well traditional statistics. You will often see someone claim some great breakthrough in using neural networks for financial analysis; check to see if the author compares this results to traditional statistical analysis—if not then (s)he probably has not stumbled onto anything significantly (in the statistical sense) new. I have found, especially when it comes to my own money, that there is no substitute for learning difficult subjects like statistics. Nothing useful for such complicated activities such as financial markets should be easy to learn or apply, as is claimed for neural network."

## 4.4 Summary and Conclusions

Theoretical and empirical comparisons between ANNs and STMs has been presented. Theoretically Neural networks seem to be more powerful, capable of learning difficult problems. However, empirically the evidence does not yet show conclusive results on the

question of whether ANN is superior or not. The findings drawn by most studies are based on particular problem domains rather than on general conditions. To achieve a better understanding of their similarities and differences, one should go beyond just case studies to establish their effectivenesses.

As Hill et al. [1994] pointed out: theory-based research should identify problem characteristics that predict when ANNs will forecast better than statistical models; theory-based research should identify which input variable characteristics predict when ANN will improve model estimation; theory-based research should identify when this advantage will give substantially improved forecasting performance.

It is believed that simulation analyses on the basis of comprehensive data conditions may provide a way to solve these problems, and to discern the potential contributions and dangers between ANNs and STMs applied to the classification problem.

In later chapters a broad range of simulated data will be generated and used to test the performance and robustness to some modelling assumptions for two statistical methods – MDA, Logit as well as for two ANN algorithms – GDR and Projection.

## **Chapter Five**

### **MULTICOLLINEARITY AND FACTOR ANALYSIS**

#### **5.1 Introduction**

The main goal of this study is to evaluate the predictive abilities for four different discriminating techniques under comprehensive data situations in a bankruptcy prediction model. This evaluation is based on both an analysis of simulated data and a real data set. Before we develop the research design, some issues should be clarified in advance. The important issues to be discussed are: How should input variables (predictor variables) be generated in a simulation study? Are correlated or uncorrelated variables to be developed in a simulation study? And how do we choose the final key financial indicators in an empirical study? These issues involve multicollinearity, where some of the input variables are highly correlated; the problem in factor analysis, which is used to avoid multicollinearity; and the trade-off between these two problems. This chapter will discuss this fundamental issue, and the conclusions will provide a basis for the research methodology developed in the following chapters.

#### **5.2 Multicollinearity in Bankruptcy Prediction**

A major problem in developing bankruptcy prediction models is choosing the best combination of financial ratios, and other independent variables. The lack of a definitive theory of financial distress prediction to guide the selection of predictor variables is one common criticism of bankruptcy prediction models. The criteria used in the literature for variable selection are: popularity in practice as evidenced in texts, potential relevance based on ad hoc theory, subjective judgement and *ex post* predictive success. These

criteria typically generate a huge set of variables which need to be screened. A large set of financial ratios are likely candidates for multicollinearity problems, given that commonly cited financial ratios are merely different combinations of the same finite set of accounting measures.

Multicollinearity occurs when some of the variables are not independent different measurements, but are virtually linear combinations of each other. Such inter-relationships result in inaccurate, unstable estimates of model coefficients and estimates of their variability. In addition, the relative importance of the variables cannot be determined because several variables may be measuring the same attribute. The early research in this area recognised the issue. As Johnson [1970] observes that if two variables composing a multivariate model are collinear, then the information each adds to the model is similar, and their coefficients are assigned arbitrarily.

The assumption of mutually independent ratios necessary for multivariate discriminant analysis does not hold. The use of highly correlated multiple ratios is redundant and introduces different samples as well as generating large standard errors for these coefficients. (p.1168)

On the other hand, Horrigan [1965] pointed out that collinearity presents opportunities as well as problems in this area. Collinearity between financial ratios allows most of the information to be captured by a relatively small number of ratios. However, he cautioned that the ratios must be carefully selected to avoid multicollinearity problems. A statistical technique for creating a smaller set of uncorrelated variables is factor analysis, which has been employed in a number of studies. The purpose of this note is to explain the drawback of using factor analysis as a screening device in bankruptcy prediction.

### **5. 3 Factor Analysis**

The basic idea of the factor analysis is to describe a set of  $n$  variables in terms of hypothetical "factors" based on the interrelations of the original variables.



The early development of factor analysis was due to Charles Spearman [1904]. In the following correlation matrix he studied on the tests scores of various subjects for boys in a preparatory school, Spearman has noted that any two rows are almost proportional if the diagonals are ignored.

$$\begin{bmatrix} 1.00 & 0.83 & 0.78 & 0.70 & 0.66 & 0.63 \\ 0.83 & 1.00 & 0.67 & 0.67 & 0.65 & 0.57 \\ 0.78 & 0.67 & 1.00 & 0.64 & 0.54 & 0.51 \\ 0.70 & 0.67 & 0.64 & 1.00 & 0.45 & 0.51 \\ 0.66 & 0.65 & 0.54 & 0.45 & 1.00 & 0.40 \\ 0.63 & 0.57 & 0.51 & 0.51 & 0.40 & 1.00 \end{bmatrix}$$

From another perspective, many correlations could be accounted for by a simple model for the scores. On the basis of this feature, Spearman proposed the idea that the six test scores are all of form

$$X_i = a_i F + e_i$$

where

$X_i$  is the  $i$ th standardised score with a mean of zero and a standard deviation of one

$F$  is a factor which has mean of zero and standard deviation of one

$a_i$  is a constant

$e_i$  is the part of  $X_i$  that is specific to the  $i$ th test

Thus, the variance of  $X_i$  is given by

$$\begin{aligned} \text{Var}(X_i) &= \text{Var}(a_i F + e_i) \\ &= \text{Var}(a_i F) + \text{Var}(e_i) \\ &= a_i^2 + \text{Var}(F) + \text{Var}(e_i) \\ &= a_i^2 + \text{Var}(e_i) \\ \rightarrow 1 &= a_i^2 + \text{Var}(e_i) \end{aligned}$$

Spearman's work was historically important in developing the notion of general intelligence and IQ test. In his two-factor theory which each test result is made of two elements. One that is the "general intelligence"—the common factor to all tests, and the other factor that is specific to the test. Later, his theory was modified to a general factor analysis model

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \dots + a_{im}F_m + e_i$$

where

$X_i$  is the  $i$ th test result with mean zero and unit variance

$a_{i1}, a_{i2}, \dots, a_{im}$  are the factor loadings for the  $i$ th test

$F_1, F_2, \dots, F_m$  are  $m$  uncorrelated common factors, each with mean zero and unit variance

$e_i$  is the factor specific only to the  $i$ th test, which is uncorrelated with any of the common factors and has mean zero

With this model

$$\begin{aligned} \text{Var}(X_i) &= 1 = a_{i1}^2 \text{Var}(F_1) + a_{i2}^2 \text{Var}(F_2) + \dots + a_{im}^2 \text{Var}(F_m) + \text{Var}(e_i) \\ &= a_{i1}^2 + a_{i2}^2 + \dots + a_{im}^2 + \text{Var}(e_i) \end{aligned}$$

The  $a_{i1}^2 + a_{i2}^2 + \dots + a_{im}^2$  is the part of its variance that is related to the common factors. It is called the communality of  $X_i$ . While  $\text{Var}(e_i)$  is the part of its variance that is unrelated to the common factors, it is called the specificity of  $X_i$ .

The correlation between  $X_i$  and  $X_j$  can be established as follows

$$r_{ij} = a_{i1}a_{j1} + a_{i2}a_{j2} + \dots + a_{im}a_{jm}$$

Hence two test results can only be highly correlated if they have high loadings on the same factors [Manly, 1986]. This explains how the factor analysis trims down the variable dimensions by choosing the representative one from each factor, which groups highly correlated variables together.

In an attempt to reduce high correlations among the variables entering the final model, factor analysis is often used in bankruptcy studies to limit multicollinearity and try to still capture as much information as possible from the original financial ratios. Let  $\mathbf{X}' = (x_1, x_2, \dots, x_n)$  be a vector of  $n$  independent variables (predictor variables),  $\mathbf{V} = (\sigma_{ij})$  be its variance-covariance matrix, and  $\mathbf{C} = (\sigma_{ij} / \sqrt{\sigma_{ii}} \sqrt{\sigma_{jj}})$  be the correlation matrix, where the primed quantity indicates transpose. Let  $\mathbf{W} = (w_{ij})$  be an  $n$  by  $n$  matrix of weights (factor loadings). The question is: how are these weights to be selected to achieve the goal of having zero correlation terms (uncorrelated factors) of  $\mathbf{W}'\mathbf{X}$  in which a few of them will capture most of the variability? The answer to this is connected with a property of square matrices: they have what are called eigenvalues and eigenvectors associated with them.

That is, to find  $W$  such that  $W'CW=L$ , and  $L$  has zeros in the off-diagonal. This is achieved by maximising  $(W'CW)$  with respect to  $W$  subjected to  $WW'=I$ . Forming the Lagrangian in the normal way gives  $CW=WL$ . The matrix of weights  $W$  are formed by the columns of eigenvectors, and the diagonal terms of  $L$  are the eigenvalues associated with each eigenvector. In the terminology of multivariate statistics the eigenvectors give the factor loadings, the eigenvalue is the variance of the transformed variable, and the factors are  $W'X$ . Each of the factors has a correlation of zero with each of the other factors. On the other hand, the sum of the variance of the factors will add up to  $n$  as before, i.e., the total variance of the factors will be the same as the total variance of the original variables. However, the weights can be chosen so that the first factor has the largest variance out of the set of factors, the second one has the next largest variance, and so on. As a result, it would typically happen that a few of the factors would account for most of the variance of the original variables, making further analysis more convenient.

An orthogonal rotation technique can be used to yield a set of orthogonal factors which may be easier to interpret than the factors derived from the original. Therefore, factor analysis allows the financial researcher to distill the original variable set down to a smaller distinct and orthogonal factors which contain approximately the same amount of information, with each factor being a linear combination of the original ratios. These factors may then be used as independent variables for the model. Alternatively, the original variable most closely related to each factor may be chosen. In either case, the result is a set of independent variables which capture the information contained in the original variable set but which do not suffer from multicollinearity. In practice, the ratio which loads most heavily on a given factor is usually chosen to represent the factor. Consequently, the fewer predictor variables reduce multicollinearity by decreasing the likelihood that the variables included in the model are linear combinations of each other or that they exhibit a high degree of inter-correlation. In addition, fewer variables in the final model simplify the application of the model and the interpretation of the results. Too many ratios used in a model always have the risk of overfitting, so that it is successful in classifying the derivation sample data, but less effective in generalisation and application. What is more, the reduced variable set offered the advantage of considerable time saving when large sample size was used.

With the empirical studies in accounting and finance using financial ratios to evaluate the performance and financial condition of an entity, factor analysis has been used by many authors to overcome the effects of multicollinearity. Some representatives of this work are summarised in Table 5.3.1. In these studies the same strategy has been adopted that reduction in the set of variables was achieved by selecting one variable from each factor based on the amount of the variance accounted for by that variables. As the variables that are highly loaded onto similar information, the use of more than one of variables to represent a given factor is not necessary. Factor analysis is a seductive technique which helps selecting (mainly) uncorrelated variables, condensing information about variance into a parsimonious subset.

#### **5.4 The Dilemma in Use of Factor Analysis**

When factor analysis is applied to financial distress forecasting, there are two major problems which make its value rather limited. First, in bankruptcy prediction, the variance-covariance matrices of failing and nonfailing groups are almost always different. It is inappropriate to combine these two different variance-covariance matrices into one matrix to proceed. The literature usually draws a veil over the strategy for applying factor analysis to dissimilar data sets. The second and fatal shortcoming of factor analysis is that it may destroy useful information present in the interrelationship among variables and which plays a vital role in discriminating power. Factor analysis destroys the covariance between variables, but how variables combine together is the property that is exploited in bankruptcy classification. In other words, there is a poor guarantee that variables selected by factor analysis represent all the relevant dimensions of the subject under study, and especially the dimension of covariance between variables. One can have highly correlated variables that are very good at discriminating. In bankruptcy classification nor all multicollinearity is bad. The following cases will illustrate this dilemma.

Table 5.3.1 Summary of Representative Studies Using Factor Analysis to Extract Key Predictors for Evaluating the Financial Performance

Study	Year	Variable Space and Factor Space	Methods
Pinches and Mingo	[1973]	Reduced their data set from thirty five to seven variables	LDF <sup>1</sup>
Pinches, Mingo and Caruthers	[1973]	Identify seven factors of financial position and performance from forty-eight financial ratios	Univariate Analysis
Stevens	[1973]	Six ratios were extracted from the twenty original ratios	LDF
Libby	[1975]	Reduced a fourteen variable to five orthogonal variables	LDF
Pinches, Eubank, Mingo and Caruthers	[1975]	Reduced their data set from forty-eight to seven variables	Hierarchical
Gombola and Ketz	[1983a] [1983b]	Performed a factor analysis on a set of 40 variables. Eight factors were retained, seven of which were substantially similar to the seven factors in Pinches, Mingo and Caruthers study [1973]. The eighth factor involved cash flow variables	LDF
Mensah	[1984]	Employed a factor analysis on <i>ex post</i> samples to isolate the ratios not common to factors from both group Three variables loaded highly on factors were selected	LDF
Zavgren	[1985]	Accomplished the same seven factors used by Pinches, Mingo and Caruthers [1973]	Logit <sup>2</sup>
Gombola, Haskins, Ketz and William	[1987]	Six and seven factors were obtained from the 24 candidate ratios for different research period respectively	LDF
Erxleben and Koch	[1991]	Reduced seventy-two financial ratios to the four most relevant financial ratios by factor and stepwise discriminant procedure.	LDF vs. ANN <sup>3</sup>
Poddig	[1995]	Used a combination of factor and stepwise discriminant analysis on 45 financial ratios. The results only contained three financial ratios.	LDF vs. ANN

1 LDF denotes the linear discriminant function

2 Logit denotes the logistic regression approach

3 ANN denotes the artificial neural network

### Example 1

A binary classification problem mimicking the bankruptcy prediction with two independent variables is examined. The input data in Appendix II is composed of 100 observations from each group. These two groups are derived from bivariate normal populations with the mean vector of group 1 (assuming bankruptcy)  $\mu_1=(\mu_{11}, \mu_{12})=(5.8187, 25.4445)$ , and the variance-covariance structure  $V_1=\begin{bmatrix} 1.4914 & -1.5693 \\ -1.5693 & 1.6513 \end{bmatrix}$ . The mean vector of group 2 (assuming nonbankruptcy) is  $\mu_2=(\mu_{21}, \mu_{22})=(6.4043, 19.4115)$ , and the variance-covariance matrix is  $V_2=\begin{bmatrix} 3.6397 & -5.1996 \\ -5.1996 & 7.4280 \end{bmatrix}$ . The combined dispersion matrix is  $V=\begin{bmatrix} 2.6388 & -4.2551 \\ -4.2551 & 13.6618 \end{bmatrix}$ , and the combined correlation matrix is  $C=\begin{bmatrix} 1.0000 & -0.7087 \\ -0.7087 & 1.0000 \end{bmatrix}$ .

The two groups actually are in a nearly parallel situation plotted on  $x_1, x_2$  surface as shown in Figure 5.4.1. It is evidently observed from the Figure that the covariance of  $x_1$  and  $x_2$  is high in each group. In this case we can easily distinguish one group from the other by using the cutoff surface  $x_1 + x_2 = c$  ( $c$  is a constant). However, from a one-dimensional perspective there is no discriminating power in univariate variable either for  $x_1$  or  $x_2$  (Figure 5.4.2 and Figure 5.4.3).

If we perform a factor analysis on this data set, we obtain matrix  $W=\begin{bmatrix} 0.7071 & 0.7071 \\ -0.7071 & 0.7071 \end{bmatrix}$  and matrix  $L=\begin{bmatrix} 1.7087 & 0.0000 \\ 0.0000 & 0.2913 \end{bmatrix}$ . Since the eigenvalue explains the degree of the variance accounted by the associated factor, in this case only one factor will be retained by the minimum eigenvalue criterion (factors with eigenvalues greater than one are generally considered to be significant). Hence, the conventional strategy for data reduction would be that either  $x_1$  or  $x_2$  is selected to represent this factor. Although  $x_1$  and  $x_2$  are highly correlated, in this application this is a useful property. That is, factor analysis may generate uncorrelated variables at the huge expense of throwing away inter-correlation information among variables and the discriminatory power.

### Example 2

Another three dimensional case is given in Appendix III. The 3-D graphics of the data set are displayed in Figure 5.4.4. Obviously, the two groups can still be distinguished neatly by using  $ax_1 + bx_2 + cx_3 = d$  ( $a, b, c, d$  are constants). This means that all  $x_1, x_2, x_3$  information is needed. However, the factor analysis results suggest that only one factor is retained, and that the highest loading in this factor is  $x_1$  (Table 5.4.1). Put another way, just one variable is chosen to enter the final model. If a decision is made to follow these results, none of the discriminating techniques, whether of a linear or nonlinear method, can offer good classification accuracy because the variables selected have no explanatory or predictive ability in a model which incorporates information on all facets of the group's condition.

The real financial data in the bankruptcy prediction area, of course, can not be so well-managed as the above 2-D and 3-D examples. However, as we have experienced before, financial ratios are sometimes merely different combinations of the same finite set of accounting measures, and a high correlation between financial ratios often occurs. Let us imagine that some of the financial ratios are initially selected in order to distinguish bankruptcy from nonbankruptcy. Two or three of the ratios in the two distinct groups have high within-group covariance, and are linearly dependent (or close to being linearly dependent), as is very similar to our examples. Accordingly, the results of factor analysis will clearly lose us valuable information.

The main point of these two examples is to reveal the dilemma faced by factor analysis. When the case is extended to an  $n$ -dimensional situation, sometimes the covariances structure between independent variables, undoubtedly play a vital role in the classification problem of financial ratio analysis.

## **5.5 The Trade-off between Multicollinearity and Factor Analysis**

Despite the problem that the existence of collinearity among the variables affects the stability of the underlying parameters and makes the interpretation of the role of the several attributes difficult, the application of factor analysis may destroy the discriminating

power produced by the correlations between variables. Horrigan [1965] stated that a selection of collinear ratios which are related to a dependent variable in the same fashion would obscure and possibly worsen the results of multivariate analysis. On the other hand, collinearity of ratios could be useful if one of the ratios is not related significantly to the dependent variables. Cochran [1964] has shown, however, that seemingly insignificant or unimportant variables on a univariate basis may be very important when combined with other variables. In fact, he concluded that any negative correlation and extremely high positive correlations increase the discriminatory power of a variable set, while moderate or low positive correlations may not help much, if at all. Therefore, despite a concern for the multicollinearity problem, to totally exclude highly correlated variables just because of the belief that "multicollinearity" is harmful is an even worse solution if classification accuracy is our primary objective. The cure may be worse than the disease. Moreover, multicollinearity is not such a damaging problem, and moderate departures from the assumption of mutually uncorrelated independent variables do not significantly impair the results [Stevens, 1973]. Eisenbeis [1977] also commented that multicollinearity is a sample property that is largely an irrelevant concern in discriminant analysis except where the correlations are such that it is no longer possible to invert the dispersion matrices. Altman and Eisenbeis [1978] reached a similar conclusion that the only time that multicollinearity does matter in discriminant analysis is if it is severe enough to preclude inversion of a dispersion matrix used in calculating the coefficient. In this case the coefficients cannot be estimated. Furthermore, the multicollinearity problem is coincident with using the general linear model. When employing other discriminating techniques, such as artificial neural network (ANN) approaches with a nonlinear feature, there is no need to carry out factor analysis before testing predictive ability.

## **5.6 Summary and Conclusions**

A very large number of financial ratios and other variables have been widely used as indicators in evaluating a firm's financial status. Factor analysis is a popular technique to cope with the need for data reduction. The application of factor analysis in previous



studies, although it successfully avoids multicollinearity, was nevertheless seen to potentially lose valuable information needed to distinguish one group from another. In addition a body of research indicates that multicollinearity is not a severe problem in discriminating analysis, since, strictly speaking, it does not affect the predictive ability of the function.

In the past and recent simulation studies, which compare the effectiveness of linear discriminant analysis with other approaches such as quadratic discriminant function, linear programming models, logistic regression or newly developed artificial neural networks, researchers have usually created data with zero covariance between independent variables [Lachenbruch, Sneeringer and Revo, 1973], [Bajgier and Hill, 1982], Denton et al., 1990]. They tended to treat the data as the results after factor analysis. However, it has been shown that this kind of data is unable to cover all relevant dimensions for classification purposes. Hence the strategy in these simulation studies may be flawed. It should be stressed that, multicollinearity only occurs in the linear model. When linear models are compared with other nonlinear methods, uncorrelated variables data may obscure the advantage of the latter. It appears that factor analysis is a hazardous technique to use in order to extract the key predictors.

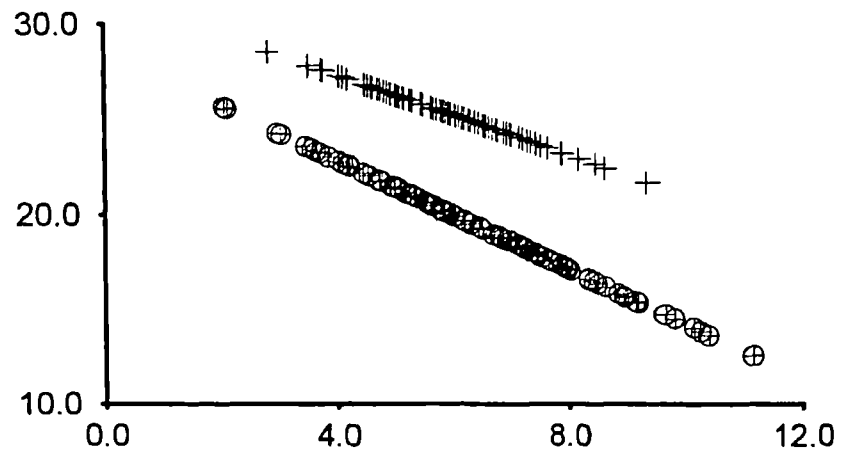


Figure 5.4.1 Scatter-Plots for Two Bivariate Normal Distributions

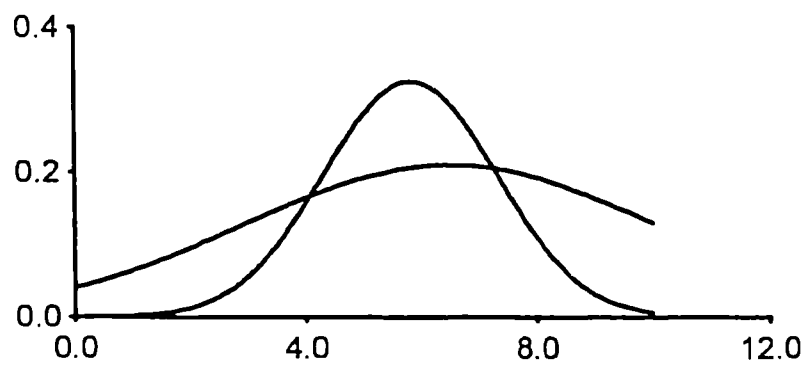


Figure 5.4.2 Univariate Distribution Plots of  $X_1$  for Two Groups

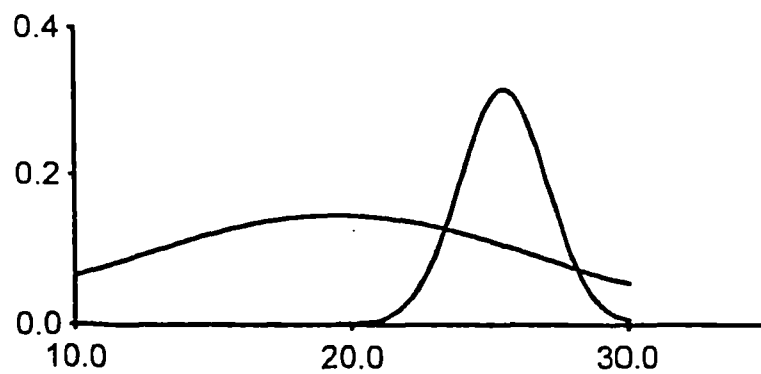
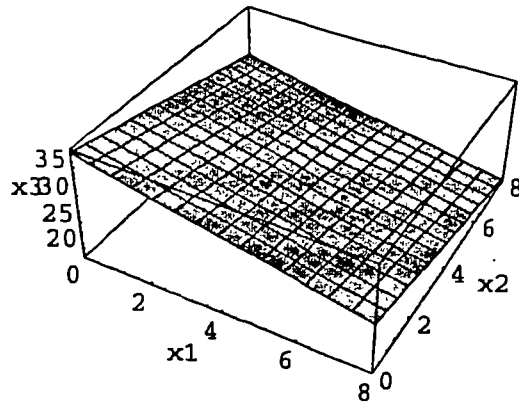
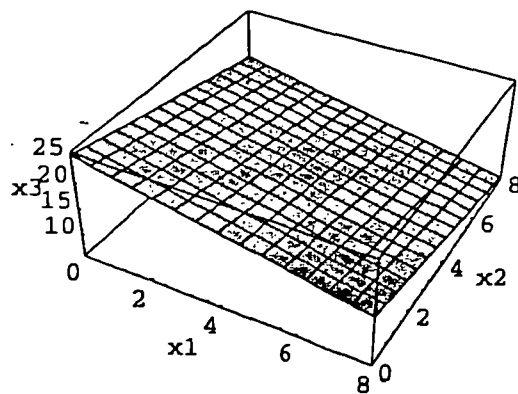


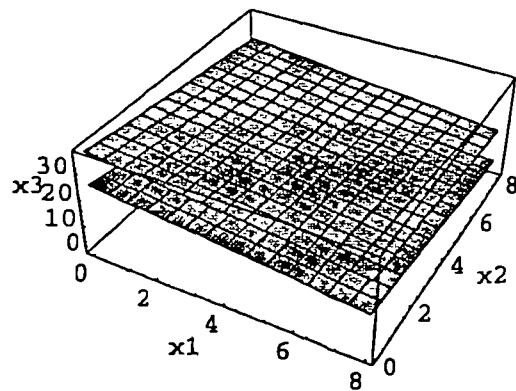
Figure 5.4.3 Univariate Distribution Plots of  $X_2$  for Two Groups



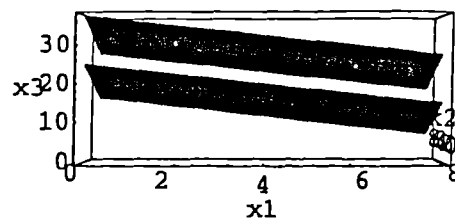
3-D Graphics for Group 1 Data



3-D Graphics for Group 2 Data



3-D Graphics for Two Groups Surface Together



The Above Graphics with a Different View Point

Figure 5.4.4 Various 3-D Graphics for Example 2 Data

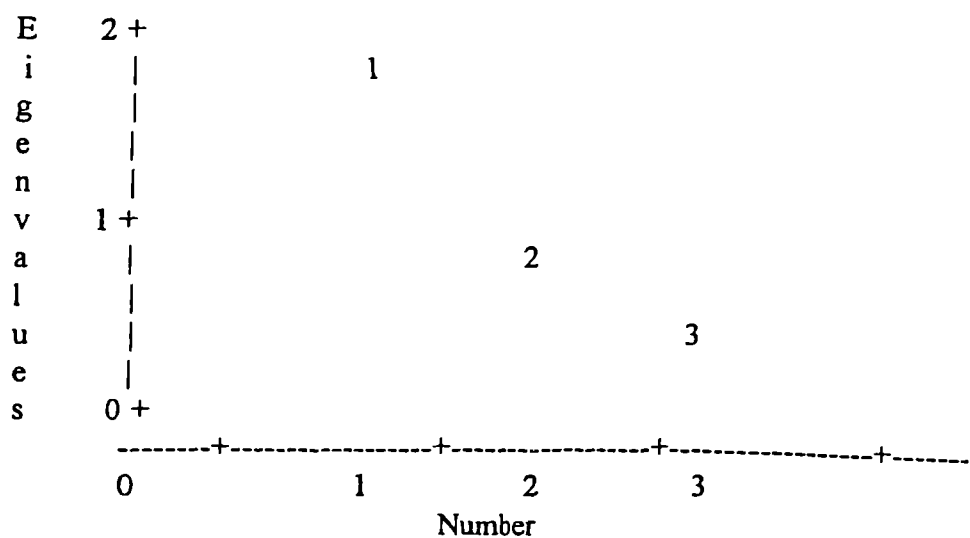
Table 5.4.1 Factor Analysis for Example 2 Data

Factor Method: Principal Components

	1	2	3
Eigenvalue	1.750726	0.797211	0.452062
Difference	0.953515	0.345149	
Proportion	0.5836	0.2657	0.1507
Cumulative	0.5836	0.8493	1.0000

1 factors will be retained by the MINEIGEN criterion.

Scree Plot of Eigenvalues



Factor Pattern

FACTOR1

R1	0.72089
R2	0.70180
R3	-0.85937

Final Community Estimates: Total = 1.750726

R1	R2	R3
0.519686	0.492526	0.738515

## Chapter Six

### METHODOLOGY OF THE SIMULATION STUDY

#### 6.1 Motivation of the Simulation Study

The theories and prior empirical studies of MDA, Logit and ANNs, as well as their comparisons in bankruptcy prediction, have been outlined and discussed in detail in previous chapters. Most studies claimed to have achieved high classification accuracy by applying the specified models to their particular empirical situations. In other words, all of the solutions have performed well when conditions favourable to the specific models are present. The researcher, therefore, can usually be assured of a suitable technique for his problem if he chooses a model which fits his situation. Although the comparative evidence related to ANNs and STMs in certain empirical cases indicated that the neural network seems to be good at solving some forecasting and classification decision problems, it is still too early to say whether these conclusions can be validated in terms of comprehensive data. No matter whether statistical methods or neural networks have shown superior performance, specific conclusions may lead to inappropriate use and invalid interpretations when applying these different techniques in other data sets.

Denton et al. [1990] explored this issue in a comparative simulation study under a variety of modelling assumptions. They compared the BP network to three other techniques: linear discriminant analysis, quadratic discriminant analysis, and linear programming with the model which maximises the sum of all distances between groups. These four techniques were compared in four cases. In all cases data was randomly generated. The cases examined were

- (1) Normal populations with equal group dispersions and a small degree of overlap
- (2) Normal populations with unequal group dispersions and a small degree of overlap
- (3) Normal populations with unequal variance-covariance matrices where the high-variance group totally overlaps the low-variance group
- (4) An identical repetition of the first case, but with one data point in the training set replaced by an outlier

Two attributes were used in input data using binary classification, with a sample of 25 observations in the training set and 75 observations in the validation set from each group. The covariance between the two attributes  $X_1$  and  $X_2$  was designed to be 0 in all four cases. That is, these two independent variables were generated independently. Each case was replicated four times.

In case 1, the results demonstrated that all four techniques achieved a good performance, but BP network obtained the best success rate (98% comparing to 93.7%, 92.7% and 93.0% respectively). When the populations were generated with unequal variance-covariance matrices, all the methods' performances deteriorated, but the neural network shows the smallest deterioration. For the total overlapping groups, linear discriminant analysis and linear programming produced success rates of under 50%; quadratic discriminant analysis improved this to 66.8%, and the neural network was close to this performance with a 65.5% success rate. In the case with an outlier, all methods were adversely affected. However, the neural network proved to be more robust than other statistical methods.

As far as we know, the study developed by Denton et al. [1990] is the only attempt to date beyond the case-study design to evaluate the classification accuracy of conventional STMs and ANNs based on MDA assumption's data conditions. No similar research exploring bankruptcy prediction has yet been published. As they admitted, although their study was useful in gaining insight into the workings of the various classification techniques, it was still quite simplistic and not extensive enough in input data generation especially for bankruptcy prediction models. The success of a technique in relation to research problems does not necessarily mean that the technique will be useful for the more complicated problems usually found in practice. In effect, the orthogonal way in which the data was generated for independent variables in their study might be inappropriate and unfair, thus leading to unreliable results of predictive ability, as we have shown in preceding chapter.

Recent empirical comparative analyses of ANNs and traditional MDA or Logit methodology that claimed the former can outperform the latter has also been severely criticised. [Trigueiros and Taffler, 1995]. Trigueiros and Taffler [1995] presented five typical recent comparative studies including Salchenberger et al. [1992]; Tam and Kiang

[1992]; Sharda and Wilson [1993]; Coats and Fant [1993]; and Rahimian et al. [1993] and stated that "none of these papers is a valid comparison made between neural network and multivariate statistical methodologies nor appropriate attention paid to or awareness demonstrated in the extant literature" (p.9). They observed that too much effort has been placed on fitting the MLP, but little attention has been paid to sample and variable selection and the application of the comparative statistical technique. Therefore, they strongly recommend

"Author should first ascertain whether poor empirical performance of conventional statistical approaches is due to their inability to deal appropriately with the complexity of the underlying relationships being studied, or rather through lack of key predictors. An equally plausible reason may be that, given the set of independent variables available, there is nothing more that can be explained independent of methodology." [Trigueiros and Taffler, 1995, p.14].

In essence, the model that predicts most accurately in this case may not have same results in other situations. Thus it may not be the best or have the highest degree of reliability. In the light of previous comparative studies, although it was found that the ANNs performed at least as well as STMs, we suspect that there are only certain conditions where this is true. There is a need for more rigorous theory-based research instead of just case studies before these techniques can be fully understood and become accepted as modelling tools. The simulation study, the first part of this thesis, aims to achieve reliable results by employing comprehensive data sets and a statistically sound design to understand how the various factors and their interactions affect the prediction performance for the different discriminating techniques of MDA, Logit, GDR and Projection algorithms.

## **6.2 Experimental Design**

In order to insure that a broad range of cases was considered in our experimental design, we generated 36 test samples of 120 in size from distinct bivariate populations pairs. Cases were limited to bivariate populations in order to facilitate the interpretation of

experimental results through graphic analysis. Such a restriction also provided a substantial degree of control in generating populations with specific types of distinctive characteristics. It should be noted, however, that our ability to generalise results unconditionally to the n-dimensional case may have been correspondingly limited. In spite of this, it is hoped that we can understand more about the differences between traditional statistical tools and artificial neural networks.

Our simulation study will be conducted in two stages. In the first stage different types of data set will be produced by varying the data distributions, the variance-covariance matrices and the relative orientation of the paired populations.

The second stage of the experiment will test the classification accuracy of each of the aforementioned techniques using various generated simulation data. Three different measures of classification accuracy will be calculated: Type I error (the error of misclassifying a failed company as a healthy company), Type II error (the error of misclassifying a healthy company as a failed company), and Overall error rates. These were provided by the specified procedure both in the training set and the testing set.

A principal objective of this simulation study is to compare the performance of different methods. Consequently, we will assume equal misclassification costs and equal prior probability. Both ANNs and STMs can be modified to reflect a user's judgement concerning particular unique problem characteristics (e.g. to assign relatively high costs to misclassifying bankruptcy as nonbankruptcy). However, in this study attempting to introduce the full range of fine-tuning possibilities for each procedure and for every sample case is judged excessively cumbersome and potentially confounding to any interpretation of test results. Accordingly, each of the approaches is implemented in the most straightforward manner, free of additional intervention.

A sample size of 120 (60+60) was selected as sufficiently large to retain the effective manipulation of the distinct population characteristics. It also represents a practically acceptable discriminant problem while at the same time providing a manageable data set that can be readily analysed.

The variables concerned in the simulation study are three: data distribution, group dispersion, and the different orientation of the bivariate variables. The levels of these three factors are stated below



- (1) Data distribution: They are the multivariate normal distribution, skewed distribution and the symmetric distribution with the existence of extreme or outlying values. This factor is qualitative.
- (2) Group dispersion. This variable is designed to indicate homogeneous and heterogeneous relationship between two groups. Equal and unequal variance-covariance matrices with the combination of high or low within-group correlation between the two attributes were developed to test the impact of variance-covariance differences.
- (3) Relative orientation of the pairs: The three orientation schemes of Figure 6.2.1 were selected to provide distinctly contrasting possibilities.

The entire experimental design concept is depicted in Figure 6.2.2. The dimensions of the experiment consisted of population distributions with three levels, population variance-covariance with four levels and population orientation with three levels. Hence, the study involved a complete factorial design with a total of 36 ( $3 \times 4 \times 3$ ) factor combinations. The dependent variables are the misclassification rates of the *MDA*, *Logit*, *GDR* and *Proj* discriminating techniques. That is, 4 techniques  $\times$  3 distribution  $\times$  4 group dispersion  $\times$  3 relative orientation situations were carried out .

To establish steady state conditions, each type of situation was replicated 20 times for training and validation data respectively. Classification accuracy of a discriminant procedure is best determined by applying the discriminant rule developed using the training sample to a validation sample. Reapplying the discriminant rule to the same data from which the rule was generated would give an unrealistically high rate of correct classification. The use of a validation sample provides a more accurate measure of the *performance of the classification rule*.

The results of both the training and validation sample data were further analysed using the Multivariate Analysis Of Variance (MANOVA) model to identify which method classify specific types of data as well as indicate whether the result is consistent between training and validation samples.

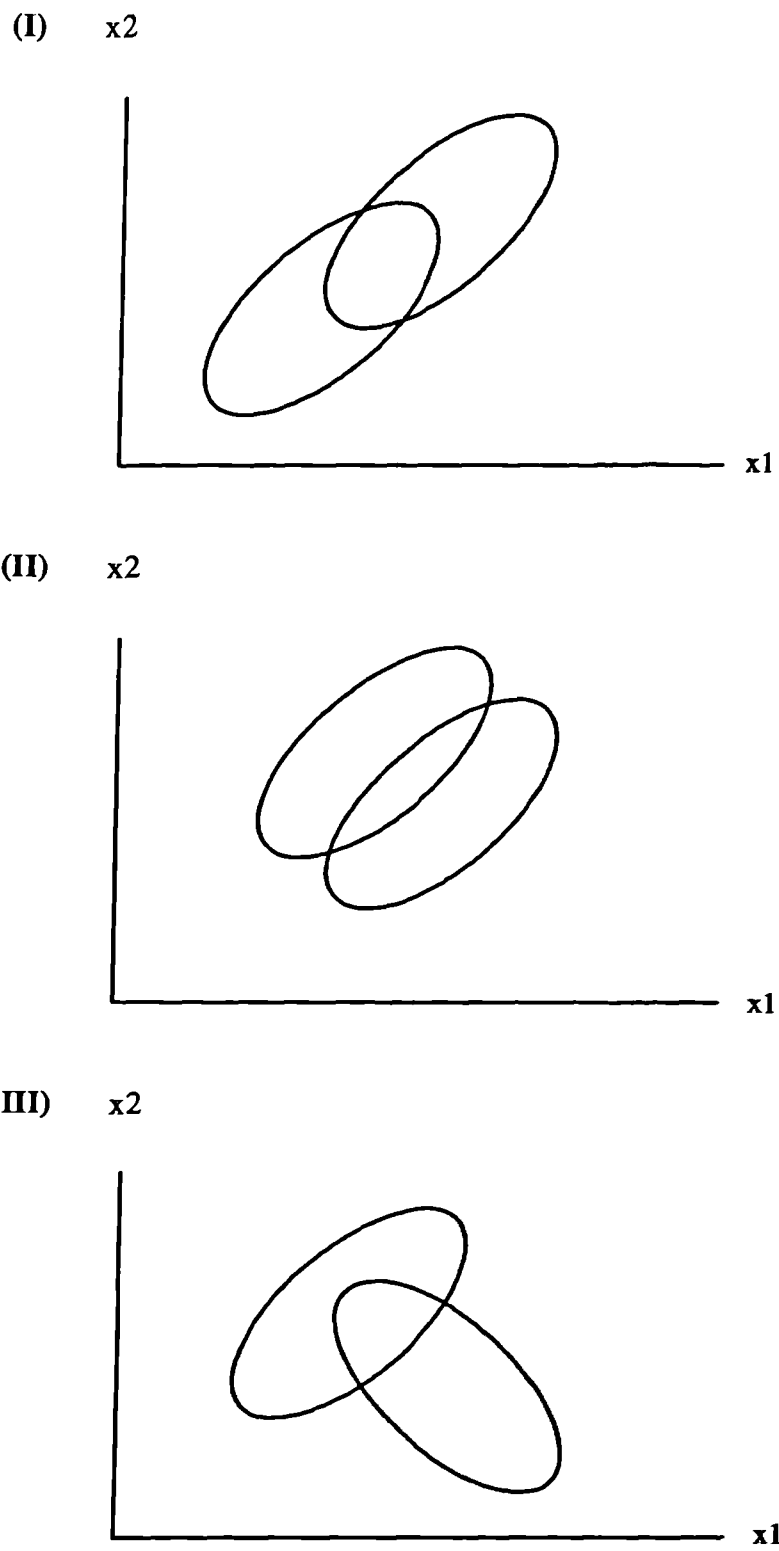


Figure 6.2.1  
Three Orientation Schemes for Bivariate Variables Experimental Design

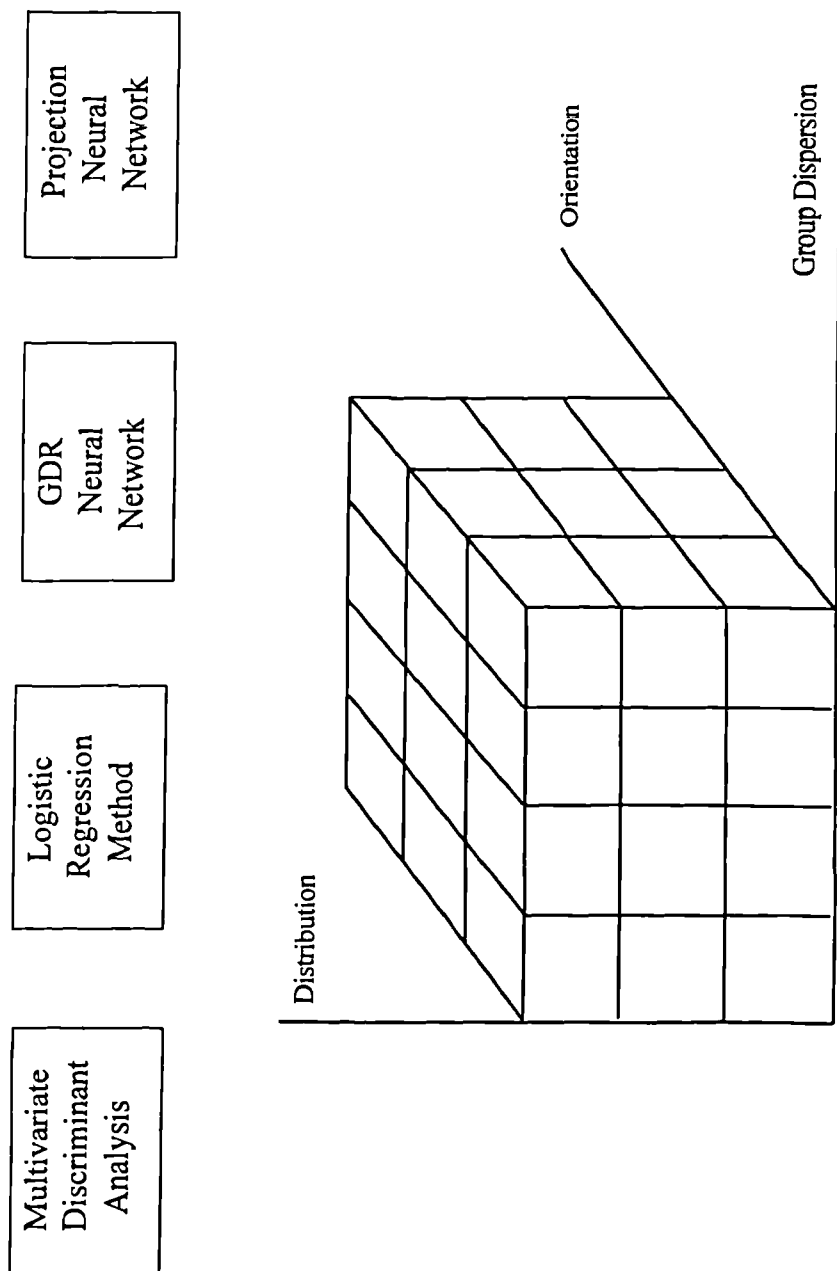


Figure 6.2.2 Experimental Design for Comparisons of Conventional Statistical Methods and Neural Network Approaches Through the Simulation Study

This study was motivated by the limited amount of research on the relative effectiveness of traditional discriminant procedures compared with the ANN approaches under a wide variety of testing conditions. Although the basic experiment is by no means an exhaustive test of the procedure's potential, we established the following goals

- (1) To promote a reasonably comprehensive classification performance comparison of the ANN approaches with conventional discriminating techniques.
- (2) To establish whether the GDR and the Projection ANN are promising procedures in predicting bankruptcy under a broad range of conditions.
- (3) To isolate those techniques that fail to produce adequate discriminant capacity and the situations in which this occurs.
- (4) To suggest special-case modifications which may correct apparent deficiencies.

There are three comparisons involved: The first comparison focuses on the prediction capabilities among the four methods over all ranges of data designs. We attempt to identify the problem characteristics when one discriminating procedure is superior to the others. The second comparison investigates the main effects of three factors on an individual approach. The third comparison tests the possible interaction effects of factors in terms of each of the four methods.

## **6.3 Data Set Generation**

An objective of this simulation is to obtain results which can be generalised and can be employed in real world data. Generalisation will be possible if the data sets generated are comparable to real data sets. To reach this goal, the choice of respective factors and their levels was based on the suggestions or the findings in previous theoretical and empirical studies. These will now be discussed.

### **6.3.1 Data Distribution**

One of the Fisher's Linear discriminant model's assumptions is multivariate normality in predictor variables. The effect of violation of this assumption has been discussed in many

previous studies [Werbos, 1974]; [Eisenbeis, 1977]; [Altman and Eisenbeis, 1978]; [Scott, 1978]; [Tollefson and Joy, 1978]; [Sheth, 1979]; [Ohlson, 1980]; [Pinches, 1980]; [Zmijewski, 1984]; [Zavgren, 1983]; [Karels and Prakash, 1987]; and [Odom and Sharda, 1990]. The outline of classification procedures in MDA has been presented in Chapter Two, and this particular procedure presumes that the distribution are multivariate normal with known parameters  $\mu_1$ ,  $\mu_2$  and known common covariance matrix  $\Sigma$ . In some applications the assumption of multivariate normality is not tenable. The discriminant criteria do not generally perform well in the absence of normality [Jobson, 1992, p.263]. This is the reason why in much research into financial statement analysis, the empirical ratio distributions are usually adjusted by trimming or transformation until the Normal model provides a reasonable approximation. However, sometimes after adjustment, the distribution of many financial ratios is still non-normal and asymmetrical. Choosing different distributions is a practical and a beneficial aspect to test. If the evidence in our simulation shows that the use of a particular discriminating method is robust to departure from normality, it would seem more straightforward to leave the data untransformed. As previous studies have shown, it is very difficult to identify the underlying distribution with a small data set. Without knowing the underlying distribution, it is not clear how data can be transformed to approximate normality nor how discordant observations can be identified [Ezzamel et al., 1987].

Most of these early empirical studies were aware of the existence of skewness but did not inquire into the reasons. Deakin's research [1976] is perhaps the most complete study related to the distributional properties of ratios. After examining the cross-sectional distribution of 11 ratios over the 1953 to 1972 period for large populations of manufacturing firms, Deakin noted that most of the ratio distributions were either highly skewed, flat, and/or dominated by outliers, and concluded that the normality assumption was generally not tenable except for the debt/total assets ratio. Barnes [1982] also reported that financial ratios are likely to be skewed rather than normally distributed. Despite the lack of clarification about this phenomenon, almost all studies suggest that the skewness is prevalent in financial ratios. If further analysis of the distribution family approach is considered, an obvious choice of family to model financial ratios is provided by Gamma distribution. Frecka and Hopwood [1983] suggested that this distribution may

provide a reasonable model for financial ratios. Their suggestion is primarily because this distribution is very general and includes the special cases the Exponential, the  $\chi^2$ , and the Normal distribution. The shape of the Gamma distribution is very versatile and it can adapt to a large number of situations, depending on the values of its parameters. Furthermore, with some conditions, after applying a transformation such as square-root transformation which was used to achieve normality in much of the literature, a Gamma random variable is approximately distributed as a Normal random variable. Because of this, Frecka and Hopwood [1983] developed the tests for outliers by assuming that the underlying distribution for financial ratios is the Gamma distribution. Also, many researchers have presumed the Gamma distributional property of selected financial ratios when dealing with outliers [Joshi, 1972]; [Sinha, 1972, 1973a, 1973b, 1973c]; [Veale and Kale, 1972]; [Mount and Kale, 1973]; [Kale, 1974, 1975]; [Lewis and Fieller, 1979]. Of particular importance is the fact that the Gamma distribution has been proved appropriate for skewed distributions in describing financial ratios [Mendenhall and Scheafer, 1973] and [Barnett and Lewis, 1978]. Consequently, in this study, we analyse the situation of skewed distribution in the context of a Gamma probability distribution model.

The other feature of financial ratio frequencies that has been reported in a number of previous studies is the existence of many extreme or outlying values. [Deakin, 1976]; [Bird and McHugh, 1977]; [Bougen and Drury, 1980]; and [Frecka and Hopwood, 1983]. Cochran [1963] pointed out that outliers can cause an increase in the sample variance and thus a decrease in the precision of parameter estimates. Denton et al. [1990] also demonstrated that the presence of outliers can seriously affect the prediction performance.

McLeay [1986b] tried to describe this feature by using the  $t$  distribution because this density function has more probability in the tail than the Normal. In McLeay's study, three profitability ratios (return on assets, return on equity and profit growth) have been successively fitted by respective plausible  $t$  distribution. The ratios used covered the published quoted and unquoted accounts of 1634 companies relating to periods in 1981 and early 1982 in the United Kingdom and Ireland. The evidence showed that for certain financial ratio frequencies, the  $t$  distribution appeared to be a good approximating model, in particular for the ratio of an operating flow. Considering that the outliers frequently

found in financial ratios cause the adverse effect in predictive performance, it is therefore important for a decision maker to be able to understand how seriously it may cause an adverse effect in different discriminant techniques and to choose a reliable technique, since sometimes the outliers can not be easily identified. Based on McLeay's study [1986b], the  $t$  distribution used in his paper was also designed in the present study to represent the condition of data with outliers.

### **6.3.2 Group Dispersion**

The second factor to be tested in this study is the dispersion of the group data or rather the homogeneity of the variance-covariance matrices. The group dispersion is of interest because many groups that are part of real data sets are differently dispersed. When the dispersion of the data across the two groups is approximately the same, the variance-covariance matrices are described as homogeneous. Under these circumstances it is appropriate to use a discriminating method such as Fisher's Linear Discriminant Function (FLDF), which pools the matrices when developing the discriminant rule.

When the dispersion of the data across the two groups is not the same, the variance-covariance matrices are referred to as heterogeneous, and the use of a discriminant method that pools or combines the matrices to develop the discriminant rule is not appropriate.

Some researchers have been interested in this issue. A study by Gilbert [1969] examined the effects of unequal variance-covariance matrices given some change in the prior group probabilities and number of variates on two groups. She reported that there was little difference in the classification performance of FLDF when a slight inequality of group dispersion exists. However, the prediction ability decreases when the data dispersion in one group is relatively high in comparison to the other group, and when the number of variates increased. Joachimsthaler and Stam [1988] undertook a comparative study. The goal of their research was to test and compare the performance of several discriminant models under the presence/absence of both multivariate normality and equality of variance-covariance matrices. Results of the Joachimsthaler and Stam study indicated that misclassification rates are a function of the heterogeneity of the variance-covariance.

When variance-covariance are not homogeneous, the QDF becomes the discriminator of choice. A subsequent study conducted by Stam and Jones [1990] tested the performance of the FLDF and the QDF against two LP methods. Six two-group data distributions were generated. Distributions with equal and unequal variance-covariance matrices were generated for normal, continuous uniform and discrete uniform distributions. Not surprisingly, results of the study indicated that the QDF perform better with heterogeneous groups. However, QDF performed poorly when the size of the training sample was small. The researchers recommended the use of Fisher's linear discriminant function if the sum of the observations for both groups of the training sample was less than 60.

According to the findings of previous studies, a departure from the equal group dispersion assumption must be one of our concerns. If, under a situation of heterogeneity of group dispersion (which is the usual case), the misclassification rates can be decreased with other conditions combined such as the data distribution; or if in practice, a particular technique is very insensitive to this assumption and needs not be strongly adhered to, it will be helpful for the researcher to control the variable associated with his/her study or choose a reliable discriminating technique.

For a two attributes  $x_1, x_2$  input design, the equality of variance-covariance matrices in two groups is defined as

$$\Sigma_1 = \begin{bmatrix} \sigma_{11}^1 & \sigma_{12}^1 \\ \sigma_{21}^1 & \sigma_{22}^1 \end{bmatrix} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{bmatrix} = \Sigma_2$$

where

$\sigma_{ii}^k$  is variance of  $x_i$  for group  $k$ .

$\sigma_{ij}^k$  is covariance between  $x_1$  and  $x_2$  for group  $k$

and  $\sigma_{ij} = \sigma_{ji}$

This means that every element in the two matrices should be equal. However, most previous simulation studies assumed the covariance  $\sigma_{ij} = 0$ , and varied the value of  $\sigma_{ii}$  in two groups when dealing with the unequal variance-covariance matrices. That is, diagonal variance-covariance matrices were selected. Probably the covariance is set to zero partly because of treating these variables as results after the factor analysis procedure, and partly because of its complexity if the variables are not orthogonal.



Since the covariance between predictor variables may be a vital factor in discriminating power, we have decided not to use orthogonal variables. In our study the following four situations have been used

$$\begin{bmatrix} \sigma_{11}^1 & \sigma_{12}^1 \\ \sigma_{21}^1 & \sigma_{22}^1 \end{bmatrix} = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 \end{bmatrix} \quad \text{where } \sigma_{ij}^k = \sigma_{ji}^k \quad k = 1, 2. \quad i = 1, 2. \quad j = 1, 2.$$

- (a)  $\sigma_{11}^1 = \sigma_{11}^2, \sigma_{22}^1 = \sigma_{22}^2, \sigma_{12}^1 = \sigma_{12}^2$ ; and the within-group correlation between  $x_1$  and  $x_2$  is high. That is, the covariance ( $x_1, x_2$ ) is high.
- (b)  $\sigma_{11}^1 = \sigma_{11}^2, \sigma_{22}^1 = \sigma_{22}^2, \sigma_{12}^1 = \sigma_{12}^2$ ; and the within-group correlation between  $x_1$  and  $x_2$  is low.
- (c)  $\sigma_{11}^1 \neq \sigma_{11}^2, \sigma_{22}^1 \neq \sigma_{22}^2, \text{ and } \sigma_{12}^1 \neq \sigma_{12}^2$ ; and the within-group correlation between  $x_1$  and  $x_2$  is high.
- (d)  $\sigma_{11}^1 \neq \sigma_{11}^2, \sigma_{22}^1 \neq \sigma_{22}^2, \sigma_{12}^1 \neq \sigma_{12}^2$ ; and the within-group correlation between  $x_1$  and  $x_2$  is low.

Generally speaking, the variance of financial ratios in bankrupt firms is bigger than those in nonbankrupt firms, which means for cases (c) and (d), the  $V_2 = \alpha V_1$ ,  $\alpha > 1$  where  $\alpha = [\alpha_1, \alpha_2]^T$ . In order to be closer to real data, the  $\alpha_1, \alpha_2$  are not set at the same number, but as at least the value 4 in a symmetric population and as the value 2 in a skewed population for examining the effect of those with distinctly unequal variance-covariance matrices. More specifically, the values of  $\alpha$  are chosen such that the groups are reasonably separated and the levels of overlap are possibly same across all conditions of the sampling experiment.

### 6.3.3 The Relative Orientation between Predictor Variables

For two predictor variables the relative orientation has three possibilities as illustrated in Figure 6.2.1. These different contrasting situations could differentiate the predictive ability. Freed and Glover [1986] conducted a study comparing the performance of three

different linear programming (LP) models with that of Fisher's linear discriminant approach. They included this perspective as a population characteristics consideration. The results indicated that there were different classification capabilities for different techniques. Even for the same technique, these three situations produced distinct performance abilities when other conditions were held as constant, including data distribution, the degree of separation between the paired populations and the similarity of respective variance-covariance matrices. In effect, consider the case (III) in Figure 6.2.1, if the mean levels of two groups are closer, this will be a real challenge for any linear discriminating model. However, for a nonlinear approach such as a neural network, there may be the possibility of a better solution. This view does need to be proved. Put another way, this factor is designed for testing the ability of individual techniques to cope with linear and nonlinear decision boundaries. Therefore, we are interested in seeing what effects will result from these four methods using the different relative orientations of the paired populations.

## 6.4 Parameter Selection and Correct Comparison

Random samples of  $n_1$  and  $n_2$  ( $n_1 = n_2 = 60$ ) were drawn from the above populations using the experimental design detailed in Table 6.4.1.

The mean values for most of cases were selected to ensure that the groups were moderately or strongly overlapped, since a small overlap would not be a challenge between the underlying techniques in our study [Denton et al., 1990].

Table 6.4.1 Experimental Design of the Simulation Study

	Normal Orientation			Skewed Orientation			Outlier Orientation		
	End	Side	Cross	End	Side	Cross	End	Side	Cross
Equal dispersion, High correlation	Aa1	Aa2	Aa3	Ba1	Ba2	Ba3	Ca1	Ca2	Ca3
Equal dispersion, Low correlation	Ab1	Ab2	Ab3	Bb1	Bb2	Bb3	Cb1	Cb2	Cb3
Unequal dispersion, High correlation	Ac1	Ac2	Ac3	Bc1	Bc2	Bc3	Cc1	Cc2	Cc3
Unequal dispersion, Low correlation	Ad1	Ad2	Ad3	Bd1	Bd2	Bd3	Cd1	Cd2	Cd3

The group overlap can be measured by the Mahalanobis distance statistic  $D^2$ . It is defined as  $D^2 = \{(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)\}^{1/2}$ . Under some conditions a different Mahalanobis distance describes the degree of group segregation. A larger value of  $D^2$  indicates that it is easier to discriminate between the two groups. In general the greater  $D^2$  for the two population, the lower is the probability of misclassification.

For our present study a summary of the parameters used in each case appears in Table 6.4.2. The scatter plots of the typical data for 36 cases are displayed individually in Figure 6.4.1-36.

Since the cases to be tested cover a broad range and control a variety of factors at the same time, it is impossible to control the same Mahalanobis distance across all conditions of the sampling experiment. For example, for set Aa2 and Aa3, we developed the same parameters except for the direction of correlation. The Mahalanobis distances were then forced to be different. In this case the different Mahalanobis distance is a blessing, since the greater or smaller  $D^2$  has already give us clues that there exists an impairment or improvement in the classification accuracy in MDA. We can then conclude that the direction of covariance between independent variables does make difference in predictive ability. However, for other cases, a different Mahalanobis distance may make a comparison between different conditions impossible. For instance, set Aa1 and set Ba1, the variances values in two groups are equal. But the mean levels which depend on parameters  $\alpha$ ,  $\beta$ , and thus depend on variance as well as the Mahalanobis distance, are not same. In this case, if the misclassification rate in set Aa1 is lower than Ba1 for a certain technique, it does not mean that this technique has a better performance in a normal population than that in a skewed population, because the degree of group segregation is not controlled as a constant in addition to different mean levels.

A similar problem has been encountered in previous research. Lachenbruch, Sneeringer and Revo [1973] undertook a study to estimate the error rate in linear and quadratic discrimination under departure from the multivariate normal distribution. In the two-population discrimination problem, they started with two multivariate normal distribution with independent components and unit variance but with mean vectors for the populations being different. In particular, the mean vector for population 1 was  $(\delta, 0, \dots, 0)$ ; for population 2 it was  $(0, 0, \dots, 0)$ . The value of  $\delta$  used were 1, 2, 3. After

log-normal transformation, the variance is 5.67, 34, 255 and 1884 for  $\delta = 0, 1, 2, 3$ , respectively. For  $\delta$  equal to 3, the covariance matrices for population 1 and population 2 for the four-dimensional lognormal cases are:

$$\Sigma_1 = \begin{bmatrix} 1884 & 0 & 0 & 0 \\ 0 & 5.67 & 0 & 0 \\ 0 & 0 & 5.67 & 0 \\ 0 & 0 & 0 & 5.67 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 5.67 & 0 & 0 & 0 \\ 0 & 5.67 & 0 & 0 \\ 0 & 0 & 5.67 & 0 \\ 0 & 0 & 0 & 5.67 \end{bmatrix}$$

These calculations show that the two populations have rather different covariance structures.

Since linear and quadratic discriminant procedures are also known to be degraded by unequal covariance structures in the populations, the design of Lachenbruch et al. [1973] confounded the effect of non-normality and covariance structure.

The main point of this example is to reveal that when one or other of the factors does not hold constant, the robustness analysis may lead to an erroneous conclusion. Therefore, for different group overlaps between cases, the sensitivity analysis of a particular factor for a technique should be carefully examined in order to avoid an incorrect comparison.

As a matter of fact it should be noted, for the present study, that even the same Mahalanobis distances do not necessarily represent the same degrees of group segregation, because if the assumptions are not valid that the  $X_i$ 's are not independent and identically distributed  $N(\mu, \Sigma)$ , then  $D^2$  is not appropriate for inference [Johnson, 1987].

However, for each case the comparative performances of four techniques will be no problem at all under the a different Mahalanobis distance. Since our primary goal is to compare the predictive performances of conventional statistical methods and newly developed neural networks, whether the group overlap is equal or not is irrelevant in comparative analysis.

## 6.5 Techniques Description

We choose the two statistical methods which appear to have been the most widely used in business failure prediction: the linear multivariate discriminant analysis (MDA) and the

logistic regression (Logit). Neither technique clearly provides substantially superior evidence to the other. Artificial neural networks are selected to compared with these two statistical techniques. The neural networks employed are GDR backpropagation and Projection algorithms.

The SAS/STAT system is used to determine the classification accuracy of the MDA and logistic regression. When performing MDA, specification statements employed with the DISCRIM procedure require the statements, METHOD = NORMAL and POOL = YES, to analyse the data. These statements prompt the use of a parametric or normal method to develop a discriminant function using a generalised square distance that pools the variance-covariance matrices of the two groups.

SAS procedure LOGIT is used to deal with the logistic regression approach. Both of the statistical methods are programmed using SAS macro processing language to reduce the trivial routine work.

The software NeuralWorks Professional II/Plus is implemented for the GDR and Projection neural network. The input layer consists of pieces of input data which describe the problem being solved. Each input node refers to a particular financial ratio (independent variables), namely, the two indicators in our present simulation study. The output layer is composed of a single response which reflects the situation's known outcome. In this study, one output node is used to denote an observation as being either healthy or distressed. The one hidden layer is selected because it has been proven that a three-layered neural network with appropriate hidden nodes can always derive a mapping from input to output to any degree of accuracy [Cybenko, 1989]; [Hornik et al., 1990]; [Hecht-Nielsen, 1989]. Therefore, the application of one hidden layer in ANNs for this study is appropriate and justified.

With respect to the number of hidden nodes, since it severely affects the classification accuracy and generalisation ability as we pointed out in Chapter Three, it should be properly determined in advance so that the performance comparison will be possible.

Here the number of hidden nodes was determined on the basis of the results of the Cascade-correlation algorithm (Cascor), which can automatically self-determine the number of hidden nodes necessary to detect all the features of the pattern. In such networks processing elements in the hidden layer are added incrementally. This cycle of

adding hidden units one by one is repeated until further addition of hidden nodes no longer reduces the forecast error. In fact Cascade learning is doing the opposite of pruning in that it is building up a network from scratch, whereas pruning starts off with a large network and prunes down its size.

In order to avoid overfitting and to achieve a reliable generalisation, training sets are used to learn optimal network structure and testing sets are used to validate whether the structure of the trained network is an optimal structure. Observing the results of some pilot studies, the optimal topology was determined for 5 hidden units by Cascor software. In parallel with the Cascor implementation, viewing a Hinton diagram of the network which can pictorially portray the significance of hidden layer outputs is the other tool to help us determine the necessary number of hidden nodes.

The optimal topology of 2-5-1 will be used across all types of simulation data. Figure 6.5.1 illustrates a diagram of the neural network structure.

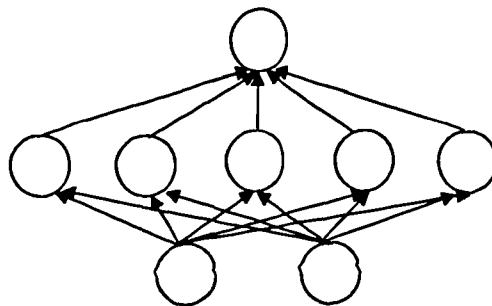


Figure 6.5.1 A Diagram of the Optimal Structure of the Neural Network in the Simulation Study

The momentum and learning coefficient are set as identical default values in each layer for all cases. The default settings of the learning rate  $\eta$  and the momentum  $\alpha$  for the BPNN are:  $\eta = 0.3$  from the input to the hidden layer;  $\eta = 0.15$  from the hidden to output layer and  $\alpha = 0.4$ . In some situations, these values may also suffice for the Projection network. However, as we showed in Chapter Three, the weights and thresholds in a Projection algorithm are initialised so that the solution is already closer to a desired output. To prevent this from causing instabilities and jumping around rather than settling on the optimum, a smaller learning rate and momentum may be appropriate, such as  $\eta = 0.01$  for all layers,  $\alpha = 0.02$ .

In addition to the learning rate and momentum, two extra parameters are required to implement the Projection algorithm. In Chapter Three, one projection from N dimensional vector into (N+1)-dimensions was expressed as

$$\mathbf{X}' = R \left( \frac{h}{\sqrt{h^2 + \mathbf{X}^2}}, \frac{\mathbf{X}}{\sqrt{h^2 + \mathbf{X}^2}} \right) \quad (6.5.1)$$

where

h: the distance between the origin of the plane (2-D) and the sphere (3-D)

R: the radius of the sphere

There are a number of such projections. In present study an alternate projection (Neural Computing, p.211), described by the following formula and which maps hyperspheres in N dimensions onto hyperspheres in N+1 dimensions, was chosen. The projection is

$$\mathbf{X}' = R \left[ \frac{2R|\mathbf{X}|}{R^2 + |\mathbf{X}|^2}, \frac{R^2 - |\mathbf{X}|^2}{R^2 + |\mathbf{X}|^2} \right] \quad \text{(see Appendix IV for mathematical derivation)} \quad (6.5.2)$$

where R is the radius of the sphere onto which the original input vectors are projected from the north pole

The last component of  $\mathbf{X}'$  in the above equation is the projected vector along the extra dimension. The remaining components lie in the original N-dimensional space.

An example of the projection for mapping circles in the original 2-D input plane onto circles in the 3-D sphere is illustrated below

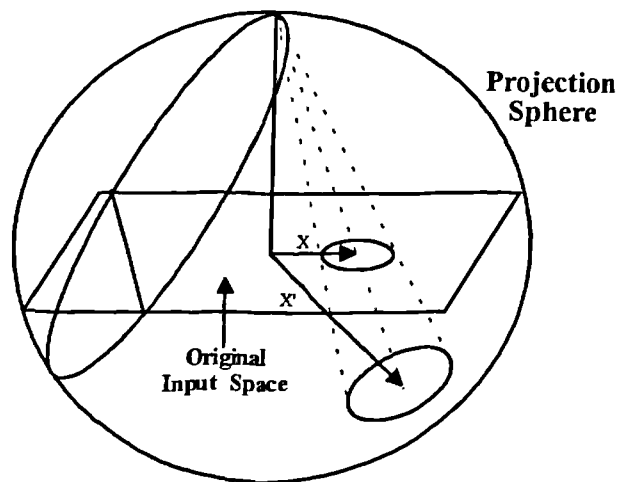


Figure 6.5.2 An Alternative Projection  
from 2-D to 3-D

The components of  $\mathbf{X}'$  given above can be easily derived by using similar triangles and the Pythagorean theorem. The  $R$  in the equation (6.5.2) represents the overall scaling of the input vectors and is the radius of the outer sphere. The magnitudes of the projected vectors  $\mathbf{X}'$  and the weights vector  $\mathbf{W}'$  are always equal to the radius  $R$ . That is,  $|\mathbf{W}'|=|\mathbf{X}'|=R$ . On the other hand, we need another parameter ( $R_0$ ) determining the radius of the inner sphere onto which the original input vectors are projected.

Thus,  $R$  and  $R_0$  are the two added parameters in the Projection algorithm. These parameters must be set properly in order for the network to achieve its optimum. Since  $R$  is an overall scaling of the input vector, it determines the steepness of the sigmoidal output of the hidden units, which has something to do with the prototype radius. As  $R$  grows, the output will approach a step function. Generally, a larger  $R$  may be useful for establishing sharp, tight prototypes with small regions of influence in an area where the output function varies rapidly and requires many prototypes. However, a large  $R$  has its danger with initial large weights leading to a very slow change during training. If  $R$  is too large, learning at the lower layer can virtually stop. Therefore, it is usually advisable to trade the sharpness of the prototypes for learning speed. After a trial and error process, the value of 6.0 is chosen as a default setting in our comprehensive data. On the other hand,  $R_0$  determines how input data is mapped to the projection hypersphere. Thus, it should be set so as to project the input vector onto a reasonable portion of the hypersphere in one higher dimension, and the separation of the input classes can be performed. If  $|\mathbf{X}| \ll R_0$ , the projection will be essentially confined to the south pole, and if  $|\mathbf{X}| \gg R_0$ , it will be confined to the north pole. Theoretically, it is best to set  $R_0$  such that all the input data will be contained within a sphere of radius  $R_0$ , that is,  $R_0$  is greater than each component of every input point:  $R_0 > X_i$  for all  $i$  and all  $\mathbf{X}$ . Nevertheless, it is tedious to apply this rule to our 1440 simulation data sets in actual practice. In the present study,  $R_0$  is set equal to the range of the inputs, and this is a good guess for the portion of hypersphere [Neural Computing, p.217]. As a consequence the default value used here is  $R_0 = \sqrt{N}$ , where  $N$  is the input dimension.

The network is fully interconnected and performs feedforward. The sigmoid transfer function is chosen for the purpose of generating all hidden layer outputs and the output layer output. 80000 epoch iterations were carried out during the network training phase.



This value was selected because it indicated that steady state root mean square error (RMS) was to be reached in those pilot studies. The RMS is determined by adding up the squares of the errors for each PE (node), dividing by the number of PEs in the output layer to obtain an average, and then taking the square root of that average. Hence the name root mean square. This RMS error is a valuable and common measure of the performance of a network during training. Additionally, the Confusion Matrix is one of the indexes to measure network performance. It indicates how to correlate the actual results of the network to the desired results. Through these two indications the appropriate value of iterations is thus determined. Furthermore, the observations are run in a random manner, as opposed to sequentially. Random presentation of training data helps the network to avoid less than optimal solutions as well as to prevent just learning the latest classification rule neglecting the prior one presented in the observations. That is, when similar data are grouped together and are presented in a sequential fashion, the network may be losing what it has learned from one end of the data set to the other. At the start it learns one set of relationships, whereas it learns a different set of relationships it moves towards the end of the data set, forgetting what it learned at first. Using a randomisation scheme can eliminate this bias.

Apart from the respective description of the usage of different techniques—STMs and ANNs—, it is very important to note the inherent philosophic differences between these two techniques. While both the statistical and artificial neural network models provide a single output for each observation, the MDA and Logit are essentially static techniques, while by contrast, GDR and Proj are dynamic. If the statistical models are to be used in practice, it is likely that they would be fixed at a certain point in time, and predictions would be made based on this fixed model. However, the network models are learning networks with the ability to learn on an incremental basis. That is, ANNs would continue to evolve as new cases are added to the network in marked contrast to the batch update procedure in statistical techniques [Weiss and Kulikowski [1991]]. Nonetheless, for comparison purposes, in our present research, the networks were taught using the training cases and then were not permitted to continue to learn as would normally be the case.

In order to run 1440 ( $36 \times 20 \times 2$ ) times cases, a user control program was created to perform training or testing networks in an off-line batch-processing mode. It allows multiple networks to be processed without intervention.

## **6.6 Cutoff Point Determination**

To compare the results obtained through MDA, Logit and two kinds of ANNs, a "benchmark" cutoff point is used, suggested by Tam and Kiang [1990] who employed a single threshold of 0.5 in their research. This cutoff point is set because we are concerned with dichotomous classification (failure vs. nonfailure). In the neural network design only a single output unit is needed. This rule matches the prediction process summarised by Judge et al. [1982]. In other words, if the posterior probability measured in MDA, the conditional probability measured in the Logit procedure, and the predicted values measured in neural networks are greater than or equal to 0.5, then the observation is classified as an event (for example, failure). If the output is less than 0.5, then the observation is classified as a non-event (nonfailure). As a matter of fact, the cutoff point is dependent upon the misclassification costs and prior probabilities. The value of 0.5 threshold is equivalent to the situation of equal misclassification costs and equal prior probabilities that we assumed in the simulation study.

## **6.7 Statistical Test of Results**

In the end, the commonly used performance indicator, Type I, Type II, and Overall error rate will be computed by averaging the 20 replications of each experiment. An experiment consists of generation of two pools of data with different population distributions, population variance-covariance matrices and population orientations, as well as testing the classification performance of each of the four methods. The results of training and validation samples will both be analysed using a multivariate analysis of variance (MANOVA). Significant MANOVA results will be followed up the appropriate and necessary univariate analyses of main and interactions effects.

### **6.7.1 MANOVA Assumptions**

MANOVA is a statistical tool which determines whether if there exist significant mean differences among groups on a combination of dependent variables, thus could have occurred by chance alone [Tabachnick and Fidell, 1989]. The dependent variables in this simulation study are MDA, Logit method and two neural network approaches. But first, the assumptions of MANOVA should be discussed. Tabachnick and Fidell [1989] indicate that the assumptions of sample size, multivariate normality of sample, homogeneity of group dispersions must be satisfied. Each of these assumptions will be considered.

#### **6.7.1.1 Sample Size and Missing Data**

In MANOVA, small sample size and incomplete data can invalidate the analysis. This requirement involves two issues: (1) to ensure that the power of the analysis is not reduced, each cell needs to have enough cases to correspond to an adequate number of degrees of freedom; (2) if more dependent variables than observations occurred, the cell would become singular, and the assumption of homogeneity of variance-covariance matrices would be untestable. In this simulation study, no data is missing. Each cell in the design contains 20 observations for each of four techniques. The number of observations is considered to be sufficient to satisfy this assumption.

#### **6.7.1.2 Multivariate Normality**

MANOVA methodology was developed on the basis of multivariate normal distribution. In other words, this assumption requires that each of the dependent variables meets the normal distribution. However, Tabachnick and Fidell [1989] have demonstrated that if the sample size of each cell provides a minimum 20 degrees of freedom for error in the univariate case, this approach will be robust to violations of this assumption. In this simulation study, the sample size of 20 observations for each of the four dependent variables produces 684 ( $3 \times 4 \times 3 \times 20 - 36$ ) degrees of freedom per cell. Therefore, the data is robust to the violation of multivariate normality.

### **6.7.1.3 Homogeneity of Variance-Covariance Matrices**

This assumption requires that the data can be pooled to create a single estimate of error under the truth of sampling data from the same population variance-covariance matrix. However, when the sample sizes in each cell are equal and large, there is no need to be concerned about violating this assumption [Tabachnick and Fidell, 1989]. For each dependent variable (technique) on experimental design, the group dispersion is controlled under the same situation for each cell and the sample sizes are equal and large. Thus, these data are also robust to the violation of the assumption of equal variance-covariance. The above discussion has shown that all the assumptions of MANOVA have been satisfied, so that the procedure can proceed satisfactorily in the simulation phase.

### **6.7.2 Effect Analysis**

The MANOVA will create a new, composite dependent variable using a linear combination of the dependent variables (MDA, Logit, GDR and Projection). In this simulation, they will separate the 36 cells as much as possible for each main effect and possible interaction.

Main effects analysis suggests that the mean differences in the composite dependent variable at different levels of the independent variables among the groups are greater than would have occurred by chance when other factors are constant. An example of a significant main effect would be if the classification accuracy is affected in any method just because of the different number of independent variables involved.

Interactions are present when, holding everything else constant, changes in the composite dependent variable over levels of one independent variable depend on the level of another independent variable. For instance, does Logit method perform better than other methods when the base rate is 1:5 and the number of variables is large?

In this study, MANOVA models will be developed using each of the four prediction methods as dependent variables, and the distribution, variance-covariance relationship, sample size, the number of input variables, base rate as independent variables. The model tests for main effects as well as interaction effects. The test statistic for the MANOVA will be Wilk's Lambda's and the level of significance will be 0.01. If the test indicates that

there are any significant main or interaction effects, then univariate analyses will be followed up. In order to avoid an inflation of Type I error, a Bonferroni procedure will be calculated a per comparison alpha. The family-wise alpha error is also set at 0.05.

## 6.8 Summary and Conclusions

The main point developed in this chapter has been a discussion of the research design, the selection of variables, data generation and the results analysis for a simulation study. The whole framework can be synthesised the following steps

- (1) Reviewing previous research in order to find out the relevant factors affecting the prediction capability of bankruptcy prediction models
- (2) Developing the bivariate-population experimental design with three factors: data distribution, group dispersion and orientation scheme
- (3) Designing the number of levels of individual factors and the respective parameters based on the suggestions of previous studies
- (4) Generating a 3 x 4 x 3 factorial design by arranging all possible combinations of the preceding three factors
- (5) Writing the SAS programs in order to generate the data sets. Each data condition is replicated 40 times equally divided into training and testing samples
- (6) Running the programmes to generate all 1440 data sets
- (7) Doing lots of pilot studies in order to tune the parameters selected until underlying population characteristics are satisfactorily achieved
- (8) Determining the optimal network architecture, the number of epoch and other relevant parameters in BP and Projection networks
- (9) Applying the four techniques – MDA, Logit, GDR and Projection to the data sets for testing the classification accuracy in learning and generalisation ability
- (10) Using the MANOVA model to analyse the main or interaction effects of factors among the four techniques
- (11) Investigating, comparing and explaining the results from the statistical evidence

The statistical results and their analyses will be provided in the next chapter.

Table 6.4.2 Parameters of Simulation Data

No	Case	Dis	S	Orien	Group 1 Data				Group 2 Data			
					Mean1	Mean2	Var1	Var2	Mean1	Mean2	Var1	Var2
1	Aa1	N	a	I	4	7	2	8	2	3	2	8
2	Aa2	N	a	II	3	6	2	8	4.5	4	2	8
3	Aa3	N	a	III	3	6	2	8	4.5	4	2	8
4	Ab1	N	b	I	4	7	2	8	2	3	2	8
5	Ab2	N	b	II	3	6	2	8	4.5	4	2	8
6	Ab3	N	b	III	3	6	2	8	4.5	4	2	8
7	Ac1	N	c	I	4	7	2	5	2	3	8	25
8	Ac2	N	c	II	3	6	2	5	4.5	4	8	25
9	Ac3	N	c	III	3	6	2	5	4.5	4	8	25
10	Ad1	N	d	I	4	7	2	5	2	3	8	25
11	Ad2	N	d	II	3	6	2	5	4.5	4	8	25
12	Ad3	N	d	III	3	6	2	5	4.5	4	8	25
No	Case	Dis	S	Orien	$\alpha_1$	$\beta_1$	$\alpha_2$	$\beta_2$	$\alpha_1$	$\beta_1$	$\alpha_2$	$\beta_2$
13	Ba1	S	a	I	2	1	4	sqrt(2)	1	sqrt(2)	2	2
14	Ba2	S	a	II	2	sqrt(2)	4	sqrt(2)	2	1	1	sqrt(2)
15	Ba3	S	a	III	1	sqrt(2)	4	sqrt(2)	2	1	1	sqrt(2)
16	Bb1	S	b	I	2	1	4	sqrt(2)	1	sqrt(2)	2	2
17	Bb2	S	b	II	1	sqrt(2)	4	sqrt(2)	2	1	1	sqrt(2)
18	Bb3	S	b	III	1	sqrt(2)	4	sqrt(2)	2	1	1	sqrt(2)
19	Bc1	S	c	I	2	1	4	1	2	sqrt(5)	4	sqrt(5)
20	Bc2	S	c	II	2	1	4	1	2	sqrt(5)	1	sqrt(8)
21	Bc3	S	c	III	2	1	4	1	2	sqrt(5)	1	sqrt(8)
22	Bd1	S	d	I	2	1	4	1	2	sqrt(5)	4	sqrt(5)
23	Bd2	S	d	II	2	1	4	1	2	sqrt(52)	1	sqrt(8)
24	Bd3	S	d	III	2	1	4	1	2	sqrt(52)	1	sqrt(8)
No	Case	Dis	S	Orien	Mean1	Mean2	Var1	Var2	Mean1	Mean2	Var1	Var2
25	Ca1	O	a	I	4	7	2	8	2	3	2	8
26	Ca2	O	a	II	3	6	2	8	4.5	4	2	8
27	Ca3	O	a	III	3	6	2	8	4.5	4	2	8
28	Cb1	O	b	I	4	7	2	8	2	3	2	8
29	Cb2	O	b	II	3	6	2	8	4.5	4	2	8
30	Cb3	O	b	III	3	6	2	8	4.5	4	2	8
31	Cc1	O	c	I	4	7	2	5	2	3	8	25
32	Cc2	O	c	II	3	6	2	5	4.5	4	8	25
33	Cc3	O	c	III	3	6	2	5	4.5	4	8	25
34	Cd1	O	d	I	4	7	2	5	2	3	8	25
35	Cd2	O	d	II	3	6	2	5	4.5	4	8	25
36	Cd3	O	d	III	3	6	2	5	4.5	4	8	25

The mean and variance values in skewed data (Gamma distribution) are  $\alpha\beta$  and  $\alpha^2\beta$

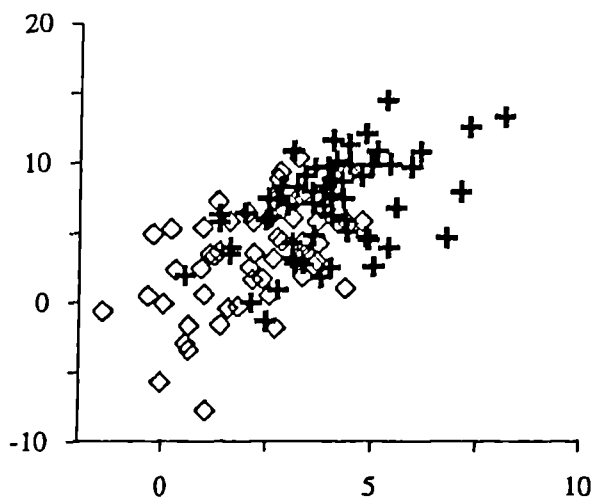


Figure 6.4.1 Typical Aa1 Data Plot

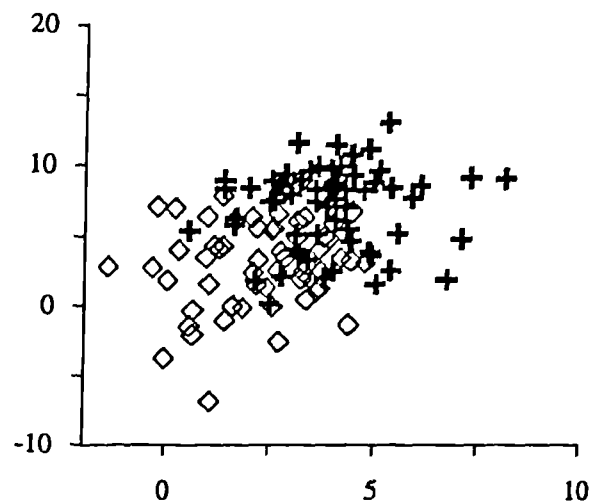


Figure 6.4.4 Typical Ab1 Data Plot

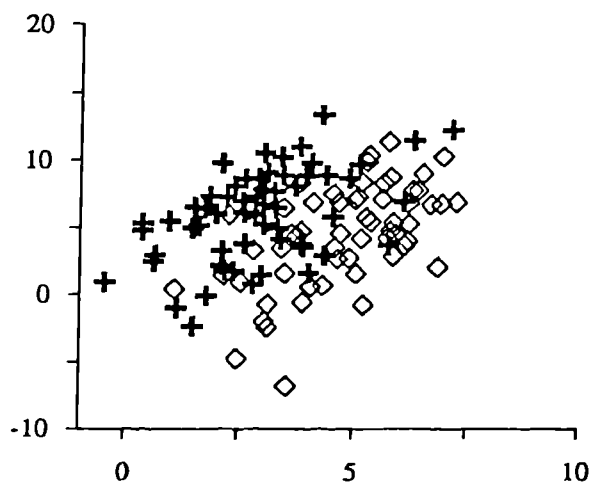


Figure 6.4.2 Typical Aa2 Data Plot

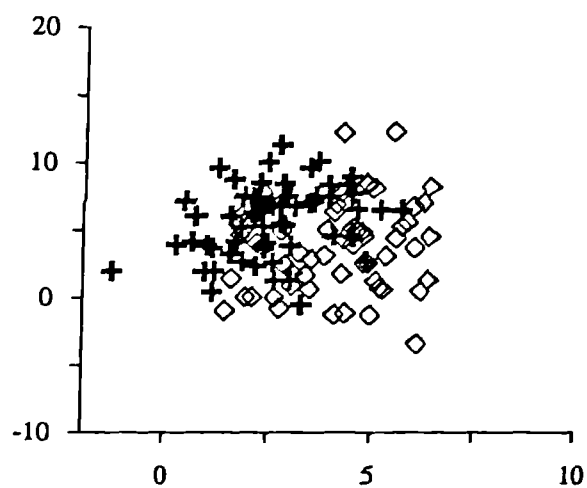


Figure 6.4.5 Typical Ab2 Data Plot

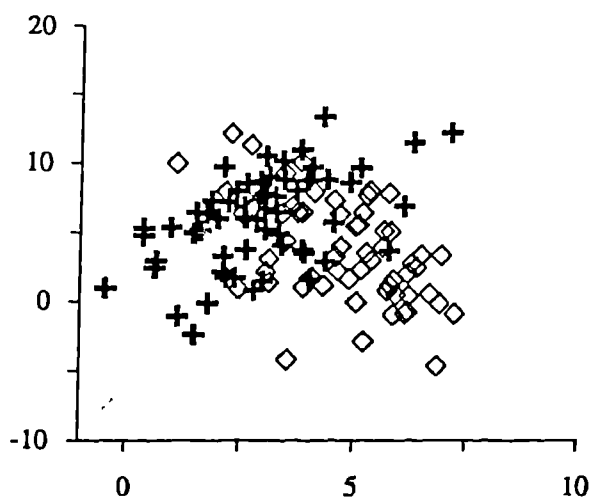


Figure 6.4.3 Typical Aa3 Data Plot

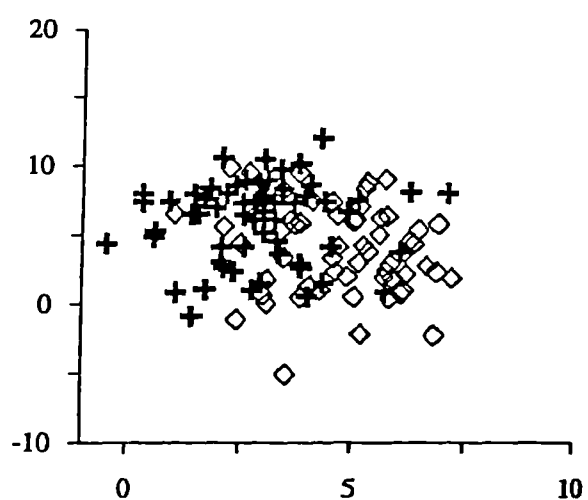


Figure 6.4.6 Typical Ab3 Data Plot

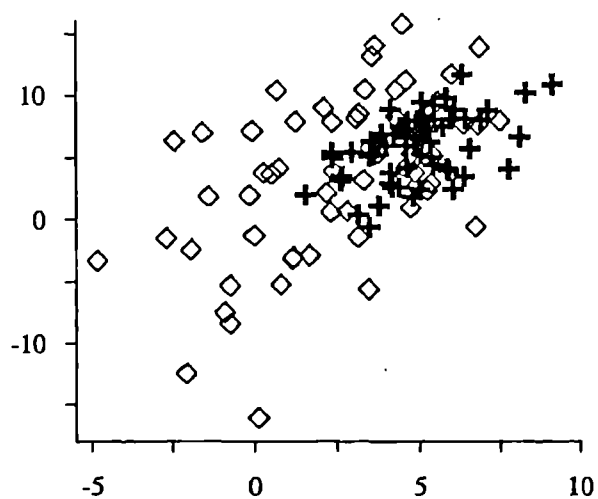


Figure 6.4.7 Typical Ac1 Data Plot

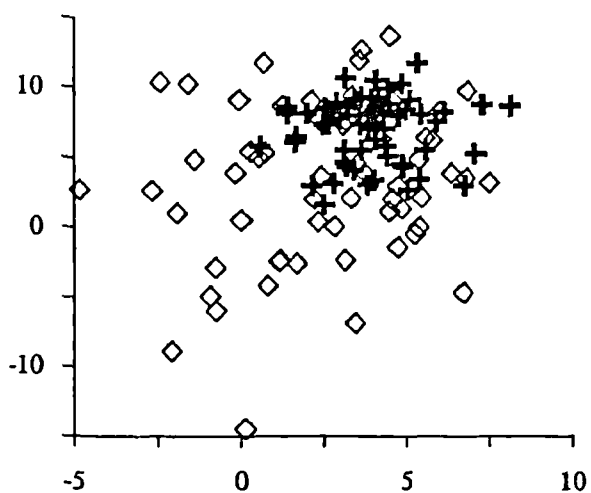


Figure 6.4.10 Typical Ad1 Data Plot

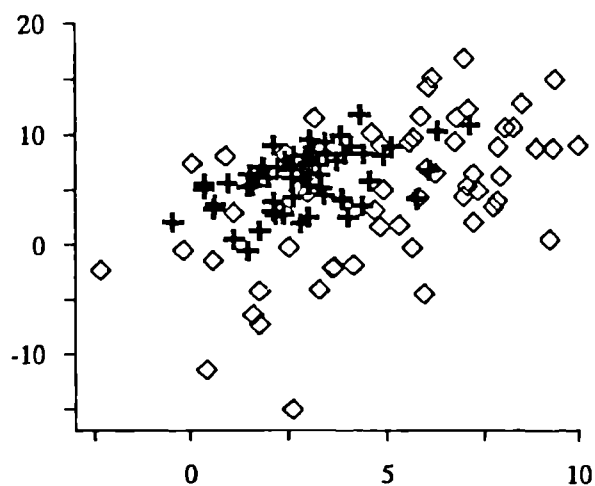


Figure 6.4.8 Typical Ac2 Data Plot

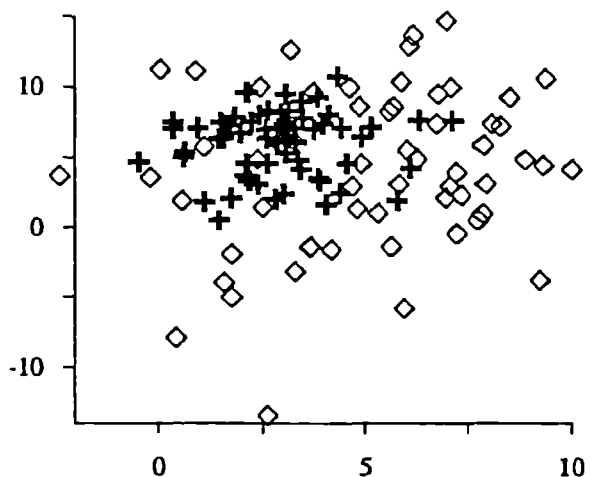


Figure 6.4.11 Typical Ad2 Data Plot

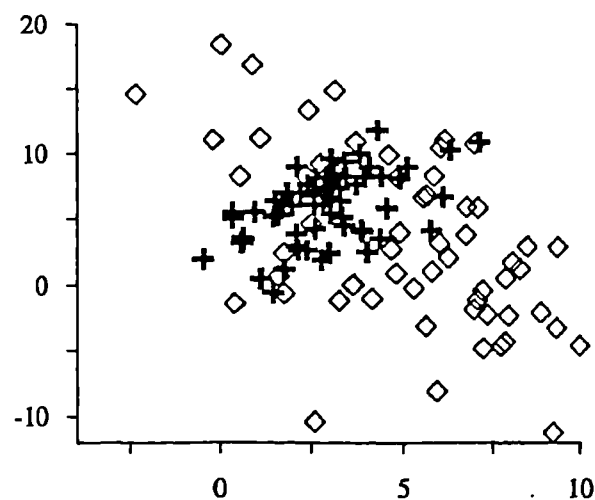


Figure 6.4.9 Typical Ac3 Data Plot

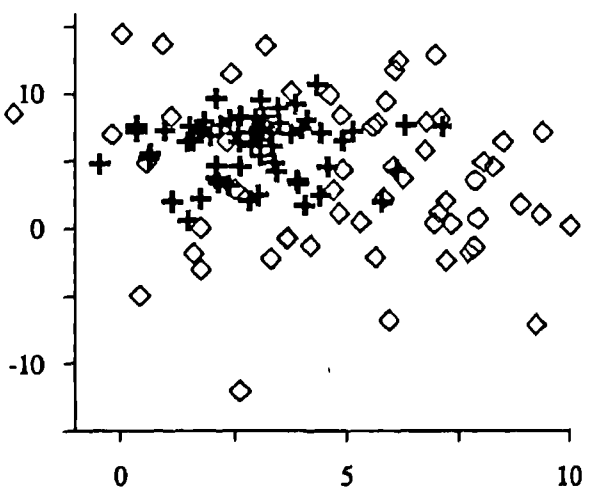


Figure 6.4.12 Typical Ad3 Data Plot



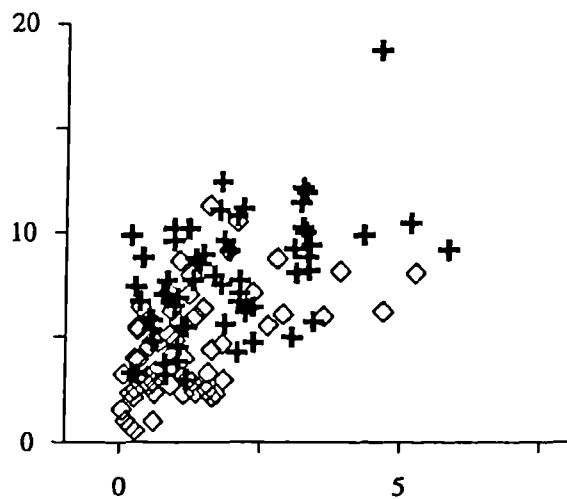


Figure 6.4.13 Typical Ba1 Data Plot

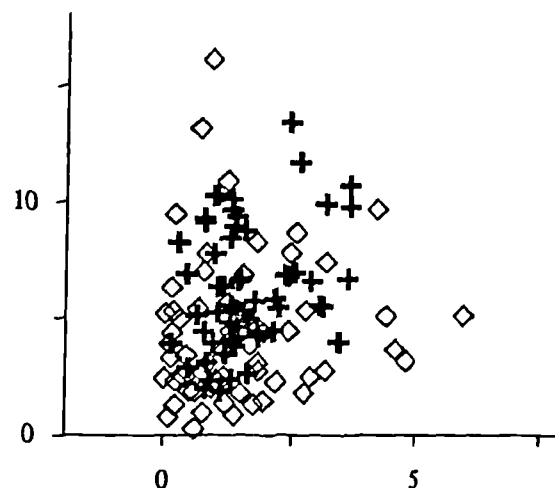


Figure 6.4.16 Typical Bb1 Data Plot

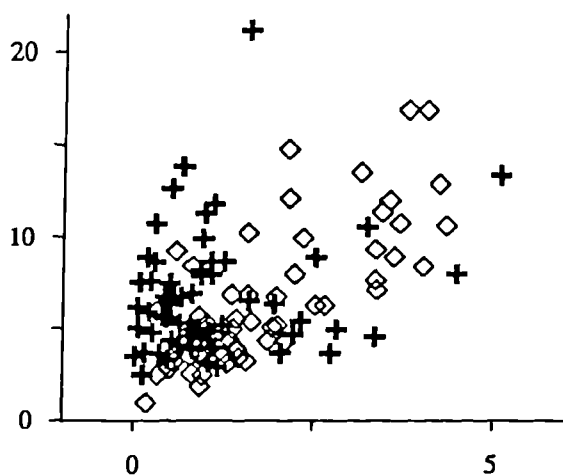


Figure 6.4.14 Typical Ba2 Data Plot

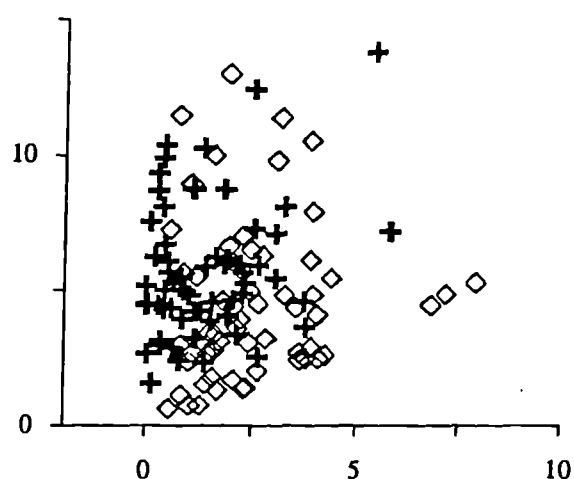


Figure 6.4.17 Typical Bb2 Data Plot

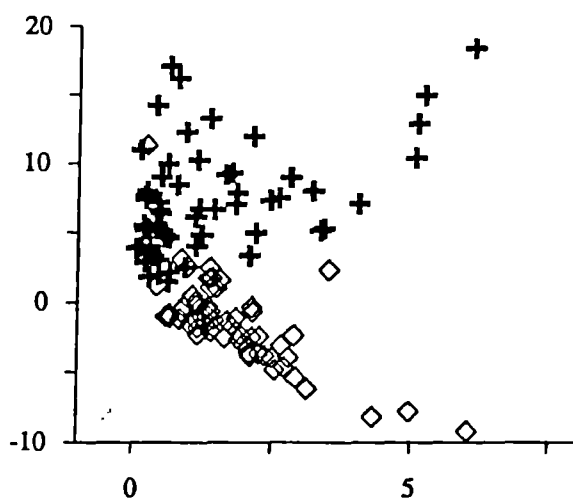


Figure 6.4.15 Typical Ba3 Data Plot

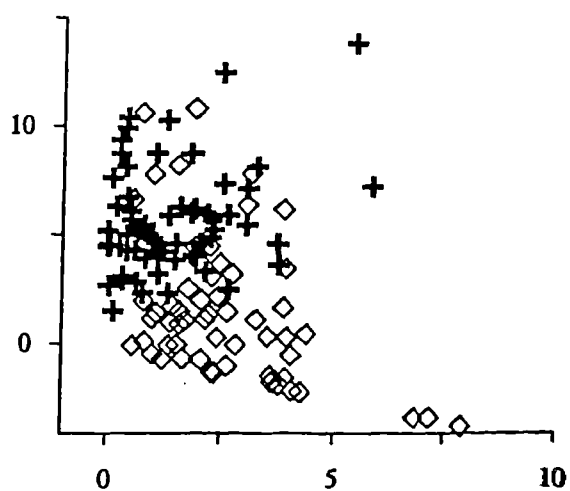


Figure 6.4.18 Typical Bb3 Data Plot

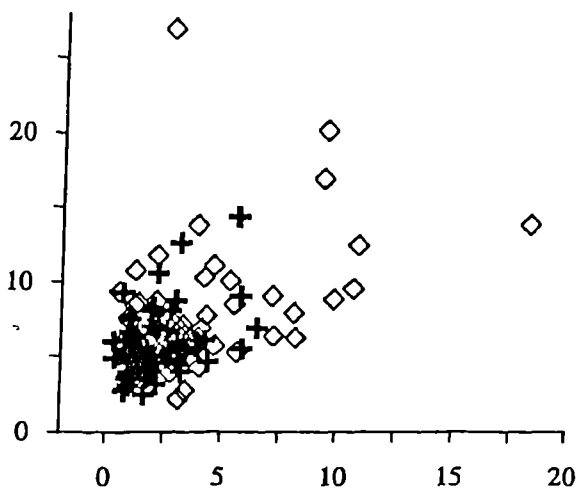


Figure 6.4.19 Typical Bc1 Data Plot

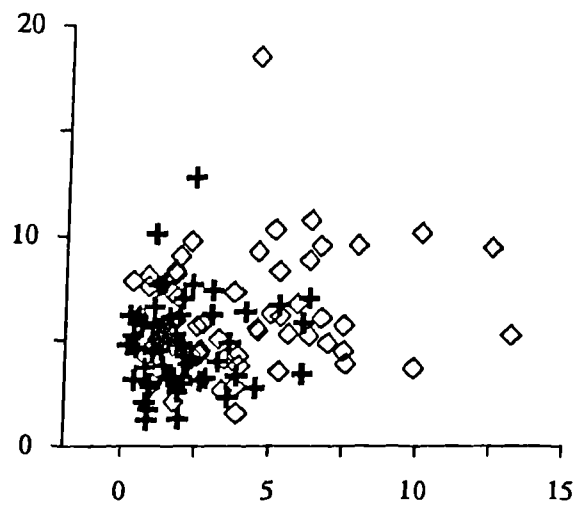


Figure 6.4.22 Typical Bd1 Data Plot

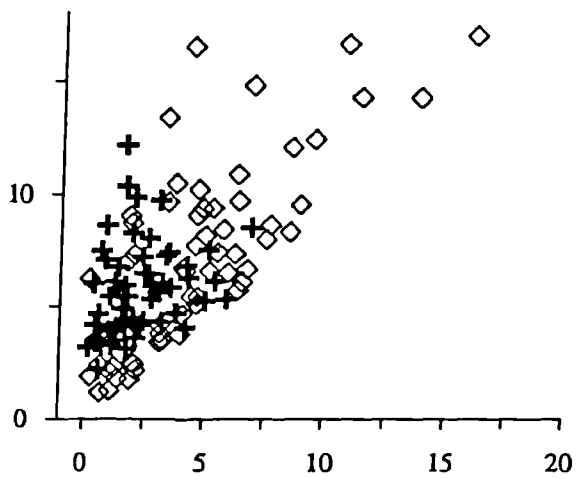


Figure 6.4.20 Typical Bc2 Data Plot

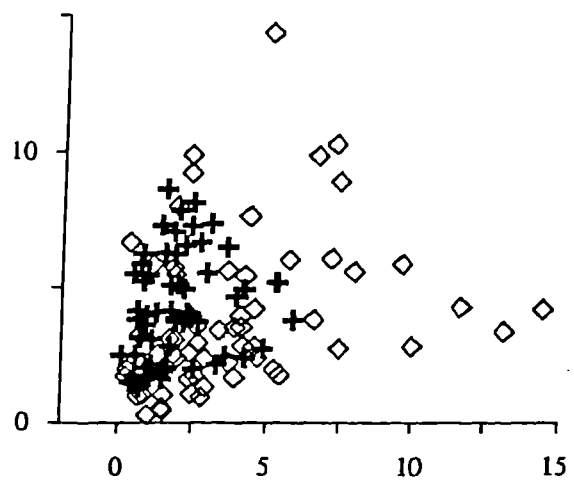


Figure 6.4.23 Typical Bd2 Data Plot

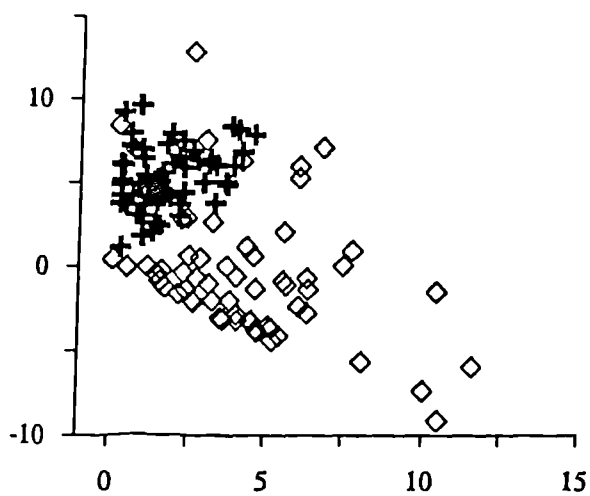


Figure 6.4.21 Typical Bc3 Data Plot

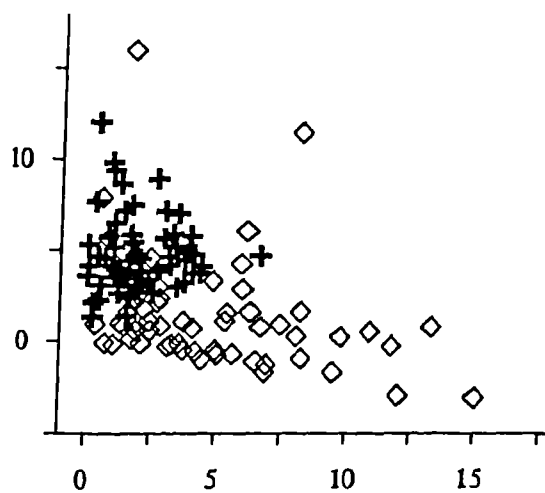


Figure 6.4.24 Typical Bd3 Data Plot

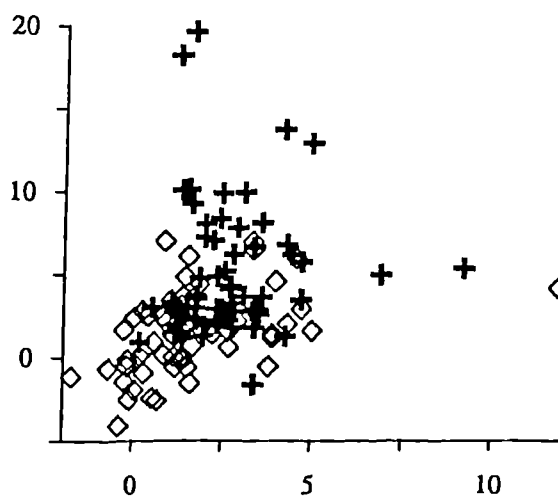


Figure 6.4.25 Typical Ca1 Data Plot

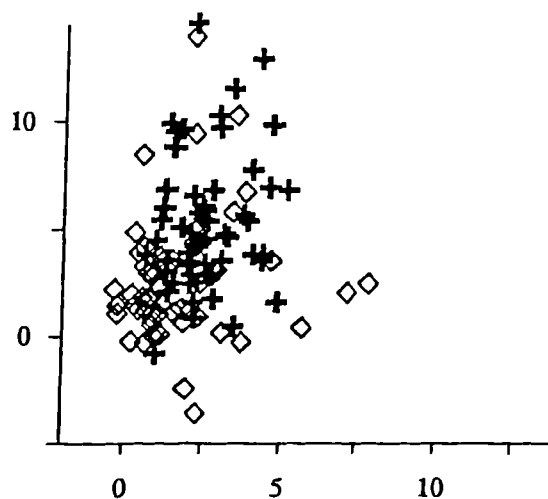


Figure 6.4.28 Typical Cb1 Data Plot

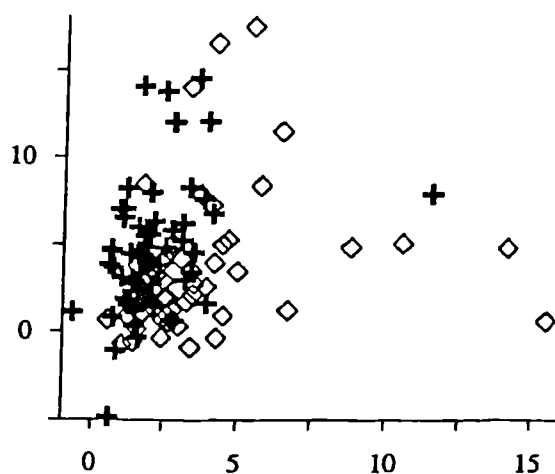


Figure 6.4.26 Typical Ca2 Data Plot

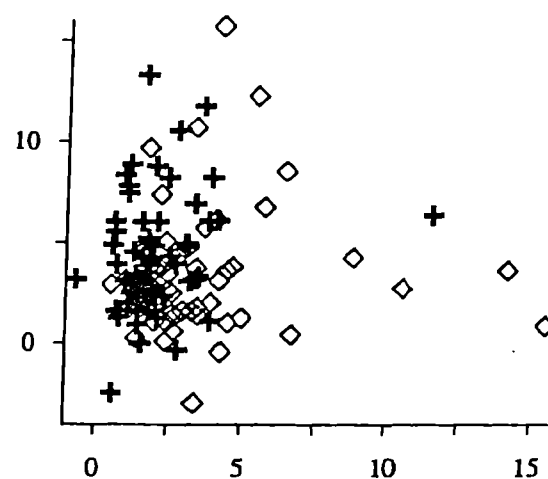


Figure 6.4.29 Typical Cb2 Data Plot

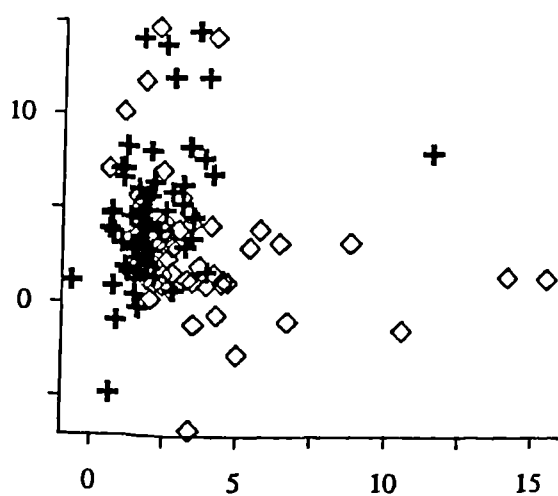


Figure 6.4.27 Typical Ca3 Data Plot

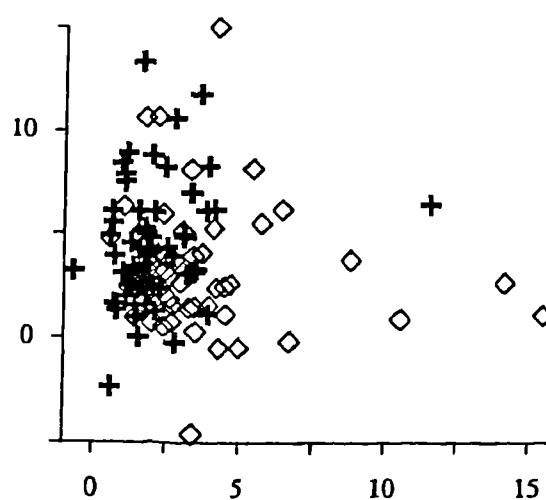


Figure 6.4.30 Typical Cb3 Data Plot

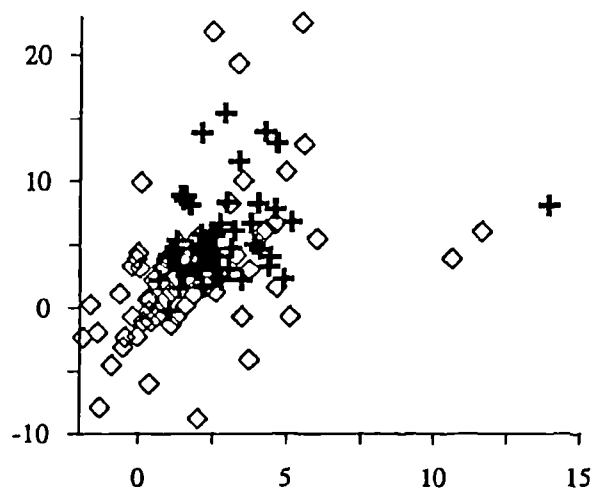


Figure 6.4.31 Typical Cc1 Data Plot

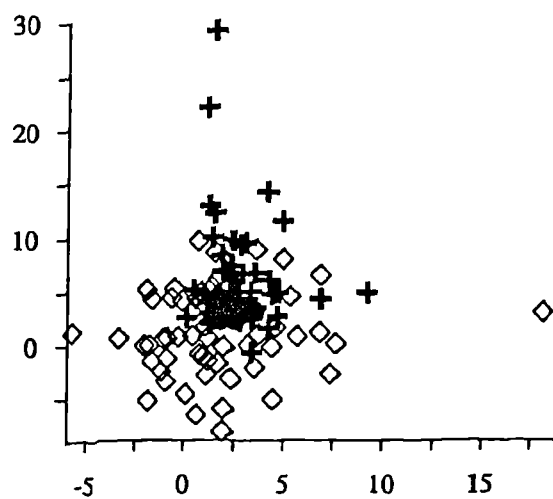


Figure 6.4.34 Typical Cd1 Data Plot

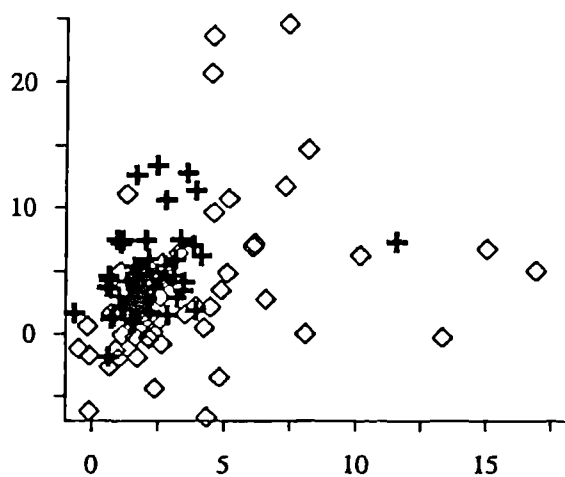


Figure 6.4.32 Typical Cc2 Data Plot

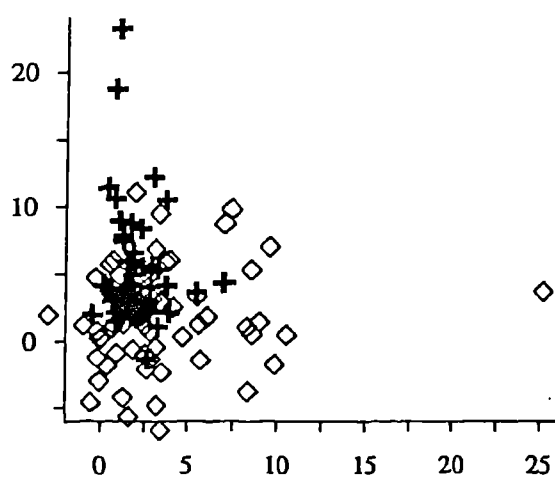


Figure 6.43.5 Typical Cd2 Data Plot

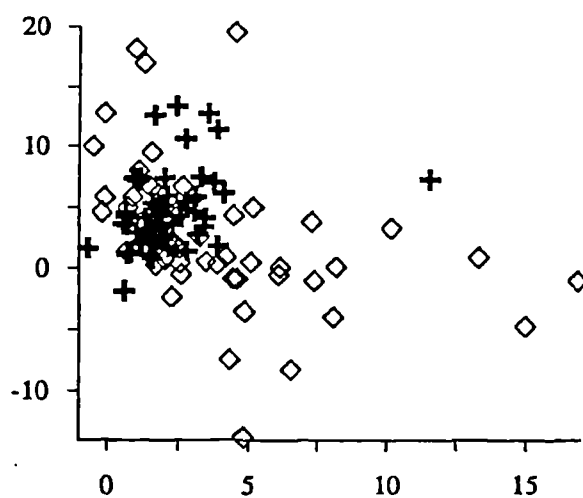


Figure 6.4.33 Typical Cc3 Data Plot

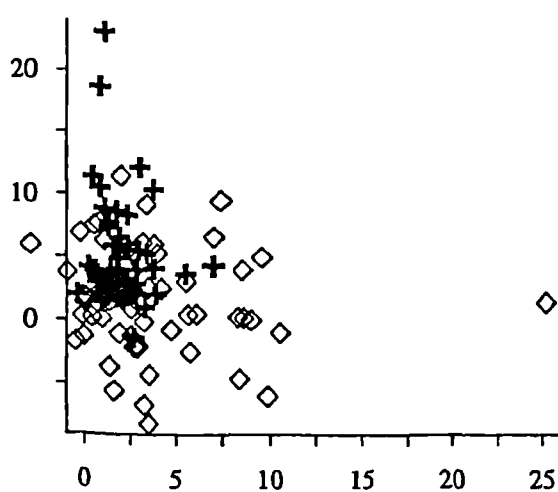


Figure 6.4.36 Typical Cd3 Data Plot

## **Chapter Seven**

### **RESULTS OF THE SIMULATION STUDY**

#### **7.1 Introduction**

This chapter reports and analyses the results of simulation study through graphics and statistical tools. The main statistical tool used to test the hypotheses is a Multivariate Analysis of Variance (MANOVA). Significant MANOVA results will be interpreted with the appropriate univariate analyses of interaction and main effects. Similarly, significant univariate analysis of variance (ANOVA) will be followed by a multiple comparison process. In addition to the introduction, this chapter is divided into four more parts. The second part describes the hypotheses tested in the simulation study. The third part discusses the outcomes on training data for four discriminating techniques. The fourth part analyses the results of the testing data. The final section offers a conclusion and summary of the discussions.

#### **7.2 The Relevant Hypotheses in the Simulation Study**

In comparing the performances of different techniques, it is very important to gather outcomes for both training and testing samples. The performance of the discriminating methods on the testing samples is of primary importance because it is the best indicator of a particular method's ability to classify new observations into groups. However, the discriminating weights or coefficients of the discriminant functions used to classify testing samples are developed using the training sample data. Thus, to comment on the superiority and inferiority; the similarities and differences of the performance of a discriminating method's ability to classify known and future observations into groups, it is necessary to analyse the results from both the training and testing sample data sets.

Based on the discussion in Chapter Six, there are three sets of comparisons. The first comparisons test the classification performance competition for some specific data conditions among four alternative techniques. Hypotheses  $H_1$  to  $H_8$  presented in Chapter One represent these comparisons.

The second comparisons involve testing the significance of the main effects of the three factors (i.e., data distribution, group dispersion and orientation) for each of the four methods. The relevant hypotheses are  $H_9$  to  $H_{11}$ .

The third comparisons test the impact of interaction effects between factors on the predictive capability for each of the four methods. The related hypotheses are stated as  $H_{12}$  to  $H_{15}$ .

The above hypotheses will be tested on both the training and testing samples respectively. Let us now start with the training sample assessment.

## **7.3 The Results and Analyses of Training Samples**

### **7.3.1 Descriptive Statistics and Graphic Analyses**

The analyses of the training data begin with an examination of the descriptive statistics. Tables 7.3.1 to 7.3.3 provide the average misclassification rates of each combination of group dispersions and orientation schemes in terms of different distributions: namely, normal distribution, skewed distribution and symmetric distribution with outliers respectively.

These misclassification rates are the result of averaging the misclassification rates of 20 replications of each experiments. In addition, Table 7.3.4 shows the pattern of Type I and Type II errors for each of the four methods.

From Table 7.3.1 to Table 7.3.3, the top rank method for each cell entry of Table 6.4.1 in the layout of Table 6.4.1 has been marked. We find that Proj and MDA are the only two techniques which feature. It tells us that the MDA in STMs and the Proj in ANNs seems more competitive in their respective area. In particular, Proj is outstandingly successful for the cases which are high skewness and in the presence of outliers.

In Table 7.3.4, the sign of + represents the situation in which Type II error is larger than Type I error. Oppositely, the sign of – means that the Type II error is lower than Type I

error. As it is presented, the MDA and Logit methods have the same patterns of Type I and Type II errors across all data conditions. Likewise, GDR and Projection algorithms always provide the same directions in these two types of misclassification except for the Cb1 and Cb3 cases. However, many differences occur between STMs and ANNs particularly in the skewed and outlier data. These results indicate that the pattern of errors between STMs and ANNs is quite different for certain cases.

In order to make a further comparison between STMs and ANNs, the differences of errors in numbers and in percentages between GDR and STMs, as well as between Proj and STMs, are calculated and reported. The misclassification rate of STMs is defined by averaging the values of MDA and Logit. Then the outcomes of GDR or Proj are subtracted by this average and displayed in Table 7.3.I in order to observe the degrees of change and relative change. The differences in absolute values or in percentages more than ten or ten percent are shadowed to represent significant differences between them.

First, as it is shown, over all ranges GDR & Proj produce better classification performances than MDA & Logit. Secondly, the most significant differences between GDR and STMs occur in situations with orientation scheme III involved (for instance, Ac3, Ad3, Bb3, Bc3, Bd3, and Cc3). However, for the Projection, apart from the cause of Orientation III, more circumstances are shown to yield significant changes. These additional circumstances include all outlier data situations. Outliers are apparently an influential element in affecting classification accuracy.

The descriptive statistics are also graphed in order to further facilitate the comparison among the four methods. Figures 7.3.1 to 7.3.3 exhibit the comparisons on predictive ability in terms of data distribution, variance-covariance relationship across groups and orientation scheme between indicators.

As Figures 7.3.1, 7.3.2 and 7.3.3 show, no matter what the data condition is, the Logit approach provides the highest Overall error rates and thus the worst classification performance. The Projection method, by contrast, has the lowest misclassifications over all 36 cases, and hence offers the best classification accuracy in the training phase among these four discriminating techniques. Furthermore, the difference in Overall errors between ANNs and STMs increases when the data exhibits extreme values. This phenomenon is quite clear if we compare the results in Figures 7.3.1 to 7.3.3. The training data used in Figure 7.3.1

and Figure 7.3.3 has been designed for the same overlap (parameters) except for the presence of some outliers in Figure 7.3.3 data (see Table 6.4.2 in Chapter Six). However, the classification power of ANNs apparently becomes greater in the cases which have outliers rather than those which are just normal distributions. This evidence reveals that ANNs are more robust to outliers than MDA & Logit.

To examine the effect of group dispersion, Figures 7.3.4 to 7.3.7 display the comparisons on four levels of variance-covariance matrices. Figures 7.3.4 and 7.3.5 represent the data situations with equal variance-covariance matrices accompanied by either high or low within-group correlation between the two indicators respectively. Figures 7.3.6 and 7.3.7 describe situations of unequal variance-covariance matrices with high or low within-group correlation. Firstly, the results indicate that when the distribution of data is multivariate normal, and when the group dispersion is identical across all predictors (i.e., the case Aa1, Aa2, Aa3 and Ab1, Ab2, Ab3), the performance of the four techniques still favours the ANN solutions. However, the difference is practically negligible.

On the other hand, the factors of different within-group correlations and between-group dispersions are found to have some impacts on MDA & Logit as well as on ANNs. Investigating Figures 7.3.4, 7.3.5, 7.3.6 and Figure 7.3.7, we note that under certain circumstances there are considerable differences in classification abilities. Unfortunately, it is unlikely simply through these figures or descriptive statistics to detect which single factor contributes the outcomes most, since the analysis was conducted on three factors simultaneously. Statistical multivariate and univariate analysis of variance may tell if the differences could have come about by chance. Hopefully, it can also provide us a clue to what are the possible causes.

With respect to the influences of different orientation schemes, direct comparison of three orientations via Figures is impossible, because some of them are not really comparable cases. The results are substantially dependent on other factors. Further, the scaling effect confounds our recognition through graphic analysis. However, from Table 7.3.I, with the use of a percentage basis, ANN solutions are clearly shown to be superior to those of STMs when the observations of two groups are strongly overlapped and the decision boundaries of overlap form like Orientation III. The underlying idea that ANNs are more compatible with the nonlinear decision set.



### 7.3.2 Change in Type I and Type II Error Rates

The last section gave us an insight into the classification performances on various data characteristics for four approaches, but only in the light of overall accuracy. In bankruptcy prediction, not only Overall error rates but also the composition of Type I and Type II errors are important, since the cost of misclassifying failed firms as nonfailed (Type I error cost) is expected to be much higher than the cost of misclassifying nonfailed firms as failed (Type II error cost). Thus, it is essential to assess the capacity of alternative techniques from a new perspective.

To explore the change of Type I and Type II error rates relative to Overall error rates, the differences in these two kinds of misclassification rates between GDR and STMs, as well as between Proj and STMs are calculated and reported. Table 7.3.II and Table 7.3.III indicate GDR and Proj results individually. The change in percentages are also illustrated by Figure 7.3.I and 7.3.II. Several interesting discoveries are summarised

- (1) Although the overall performances of GDR and Proj are better than those of STMs across all data conditions, the improvement does not necessarily result from improvement of both Type I and Type II errors. That means, the decrease in Overall errors is accompanied by the opposite increase in either Type I or Type II errors in many cases.
- (2) In GDR the learning algorithm seems to put more emphasis on reducing Type II error. 28 of 36 cases show lower Type II errors when the Overall performance is improved, whereas 17 cases indicate a decrease in Type I errors. Meanwhile, just 9 cases have reductions in both Type I and Type II errors.
- (3) In Proj the change of Type I and Type II errors is very different from that of GDR. 31 of 36 cases indicate improvement in Type II error accompanied by improvement in Overall accuracy, while 24 cases exhibit a decrease in Type I error. There are 19 cases in which both Type I and Type II errors improve. As can be seen, the Projection algorithm spreads overall improvement more equally to the two type errors. The many more cases having lower Type I error rates in Proj than in the statistical methods (24 to 12) offer evidence of the Projection's further significant contribution if the inequality of misclassification cost is a big consideration in failure forecasting.

- (4) With regard to the cases showing a worse Type I error performance in GDR, most of them occur in normal distribution. At the same time, there are no significant differences in overall accuracy between MDA and GDR under this condition. Thus, MDA is preferred to GDR in predicting bankrupt companies if the data's departure from normality is not severe, or can be transformed to a normal model.
- (5) On the contrary, the Projection approach produces higher Type I errors mostly in skewed distribution and outlier data, although it has better overall performance. In these situations the choice of Projection or MDA depends on the results intended, or the user preference.

In the light of the learning phase, ANNs produce better solutions in classification ability. In particular, the Projection algorithm overwhelmingly outperforms the other three. However, after penetrating the composition of Type I and Type II error rates, we need to be careful with the application of ANNs if decision-making involves unequal misclassification costs.

### 7.3.3 Multivariate Analysis

The graphic analyses mainly focus on the comparison of four alternative methods through the testing of Hypotheses  $H_1$  to  $H_8$ . In order to further understand the impact of all factors (i.e., distribution, group dispersion and orientation) on each of the four methods, statistical analyses will be performed. The hypotheses  $H_9$  to  $H_{15}$  are tested using MANOVA. A MANOVA is developed using each of the four discriminating methods as dependent variables, and using distribution, dispersion and orientation as independent variables. The model tests for main effects as well as interaction effects.

The main effects are assigned the variable names Dist, Disp, and Orien. Three two-way and one three-way interaction effects are possible: Dist by Disp, Dist by Orien, Disp by Orien and Dist by Disp by Orien. Wilks' Lambda is the test statistic, and significance is determined based on the multivariate F value. The results are provided in Table 7.3.5.

MANOVA results indicate that all three main effects, two-way interaction effects and one three-way interaction effect are significant at  $p < 0.001$ . These results support each of the hypotheses from  $H_9$  to  $H_{15}$ . Classification performance is affected by data distribution, the variance-covariance relationship across groups and orientation between indicators. To determine which discriminating methods are affected by main effects and interaction effects, subsequent univariate analyses are carried out.

#### 7.3.4 Univariate Analyses

A univariate analysis of variance is done for each of the dependent variables (i.e., the four discriminating methods). The purpose of this step is to trace the significant main effects and interaction effects to the discriminating method responsible for the significance. In order to avoid an inflation of Type I error, a Bonferroni procedure is used to calculate a per comparison alpha. The family-wise alpha error is set at 0.05. In order to determine a per comparison alpha error for each of the four dependent variables examined, the 0.05 was divided by 4, yielding a per comparison alpha of 0.0125. The results of the univariate analyses for each of the four methods are presented in Tables 7.3.6 to 7.3.9.

Tables 7.3.6 to 7.3.9 indicate, all three main effects (Dist, Disp, Orien) and two-way interaction effects (Dist by Disp, Dist by Orien, and Disp by Orien) as well as the three way interaction (Dist by Disp by Orien) are significant for each of the discriminating techniques at the 0.0125 level. The significant interactions will often mask the significance of the main effects, and the experimenter must usually examine the levels of one factor, say A, with levels of the other factors (say B) fixed in order to draw conclusions about the main effect of A. This idea can be illustrated graphically. Figure 7.3.III represents a plot of response data against factor B for both levels of factor A. The  $A_1$  and  $A_2$  lines are approximately parallel, indicating a lack of interaction between factors A and B. Figure 7.3.IV plots the other type of response data. Here we see that the  $A_1$  and  $A_2$  lines are not parallel. This indicates an interaction between factor A and B. Since significant interactions do not allow for main effects to be interpreted directly, the source of the interactions will be pursued first, and then the main effects will be analysed.

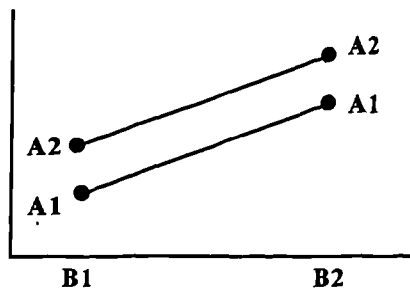


Figure 7.3.III No Interactions between A and B

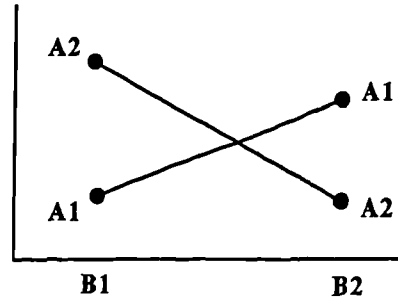


Figure 7.3.IV Interactions between A and B

### 7.3.5 Interaction Effects and Main Effects

In order to determine which effects of one independent variable (for example, Disp) vary across the levels of other variables (for example, Disp or Orien), significant interaction effects are followed up by the multiple comparison procedure. We use the Tukey approach [Keppel, 1982] for proceeding with the pairwise multiple comparisons.

Hayter [1984] gave a proof that this method controls a maximum experimental error rate when the sample size is equal. Additionally, it was found that the Tukey procedure is more powerful than the Bonferroni or Scheffe methods for pairwise comparison.

MDA method interaction effects and main effects on training data

#### 1. Three-way interaction effect (Dist by Disp by Orien)

The three-dimension picture is not easily graphed to reveal the three-way interaction effect. Table 7.3.10 provides the results of pairwise comparisons between all 36 cases. In 351 out of 630 pairwise comparisons there are significant differences in classification performance. However, as we pointed out in Chapter Six, some of these pairwise comparisons would be not meaningful if one of the other factors between comparative groups does not hold constant. Based on this rule, let us focus on the so-called correct comparisons. Firstly, the pairs (Aa1, Ca1), (Aa2, Ca2), (Aa3, Ca3), (Ab1, Cb1), (Ab2, Cb2), (Ab3, Cb3), (Ac1, Cc1), (Ac2, Cc2), (Ac3, Cc3), (Ad1, Cd1), (Ad2, Cd2) and (Ad3, Cd3) are examined, since these pairs have the same degree of overlap aside from

the presence of some extreme points in the C group. It is found that 10 out of 12 comparisons have significant differences in classification ability. These results suggest that for the MDA method, outliers appear to be an important factor in affecting predictive performance in training process.

Secondly, pairs (Aa1, Ab1), (Aa2, Ab2) and (Aa3, Ab3) are also examined because they have the same parameters apart from the difference of a high or low within-group correlation between two attributes. The results reveal that correlation between two attributes does not seem to be a crucial factor in affecting the predictive performance. In effect, after investigating all pairs with this feature, such as (Ba1, Bb1), (Ba2, Bb2), (Ba3, Bb3) and (Ca1, Cb1), (Ca2, Cb2), (Ca3, Cb3), we cannot find strong evidence that the within-group correlation between predictor variables has a significant impact on classification accuracy. The effect of this factor may be mixed up by other factors. Thus, there is a need for further evaluation by two-way interaction effects.

## 2. Two-way interaction effects

### (1) Dist by Disp effect

The results of multiple comparisons of the mean misclassification rates using the Tukey procedure are reported in Table 7.3.11. These interaction effects are also graphed in Figure 7.3.11. The picture of interaction effects suggests that the MDA method classifies data into groups more successfully when the data are of normal distribution rather than being data with outliers, no matter what structures of variance-covariance matrices they have. However, there is no clear indication that the skewed distribution would impair MDA's predictive power.

### (2) Dist by Orien effect

The results of multiple comparisons and the interaction effect of Dist by Orien are provided in Table 7.3.12 and graphed in Figure 7.3.12 respectively.

The results suggest that interactions exist between variables orientation and data distribution in the MDA method, but the source of interaction cannot be traced a position between Orientation I and II. There is more sufficient evidence that interaction occurs between Orientation III and I as well as between Orientation III and II. Moreover, whatever kind of decision boundary the data has formed, i.e.

linear (Orientation I and II) or nonlinear (Orientation III) boundaries, MDA fails to detect outliers well (the highest misclassification in C group data distribution).

### (3) Disp by Orien effect

Table 7.3.13 exhibits the multiple-comparison results and Figure 7.3.13 illustrates the effects of Disp by Orien interaction on the MDA method. The graph indicates that the misclassification rates are lower for group dispersion b than those for group dispersion d. Group dispersion b represents the data with equal variance-covariance matrices but low within-group correlation between attributes, while group dispersion d represents the data with unequal variance-covariance matrices and also low within-group correlation between attributes. These results clearly show that when the group dispersion is heterogeneous and the predictor variables have low inter-correlation, MDA does not perform as well as it does with homogenous data. However, it is noteworthy that if predictor variables are highly correlated, despite the equality of variance-covariance matrices across groups, the interaction effects with different orientations will not necessarily make the classification accuracy better. From this viewpoint, the correlation between predictor variables indeed plays a certain role in affecting predictive ability.

## 3. Main effects

All three main effects were identified as significant in univariate analysis for the MDA method. The results of the mean misclassification rates and pairwise comparisons are given in Table 7.3.14 and Table 7.3.15. These outcomes, as expected, generally demonstrate that outliers and inequality of group dispersion have an adverse impact on the classification accuracy.

## Logit method interaction effects and main effects on training data

### 1. Three-way interaction effect (Dist by Disp by Orien)

Table 7.3.16 provides all pairwise comparisons over 36 cases for Logit procedure. Examining the pairs (Aa1, Ca1), (Aa2, Ca2), (Aa3, Ca3), (Ab1, Cb1), (Ab2, Cb2), (Ab3,

Cb3), (Ac1, Cc1), (Ac2, Cc2), (Ac3, Cc3), (Ad1, Cd1), (Ad2, Cd2) and (Ad3, Cd3), it is found that the Logit formulation appears to be badly affected by extreme points in a manner similar to that of the MDA method. Furthermore, the results indicate that correlation between the two attributes provides no strong evidence influencing predictive performance when simultaneously interacted with data distribution and variable orientation.

## 2. Two-way interaction effects

### (1) Dist by Disp effect

Table 7.3.17 and Figure 7.3.14 present the multiple comparisons and interaction effects of Dist by Disp for the Logit method. Through graphic analysis, outliers are apparently shown to be a key factor in damaging classification power in the Logit method regardless of the group dispersion of the data. Moreover like MDA, the performance is better in equal group dispersion data than in an unequal group dispersion situation.

### (2) Dist by Orien effect

Table 7.3.18 reports the multiple-comparison results, and Figure 7.3.15 illustrates the effect of Dist by Orien interaction on the Logit method. The multiple comparisons indicate that the source of the interaction is traced to differences in mean misclassification rates between Orientation I and III and also between Orientation II and III. A further inspection of the graph shows that the Orientation I produces a lower misclassification than Orientation II across all different data distributions in Logit procedure, which is only slightly different from the outcomes shown in MDA. In general, the interaction between distribution and orientation has similar influences on both traditional statistical methods.

### (3) Disp by Orien effect

Even though the assumption of equal variance-covariance is not required for the Logit method, the evidence of Figure 7.3.16 still indicates that classification ability is better for homogeneous group dispersion than for heterogeneous group dispersion. This point is illustrated by achieving a lower misclassification rate in group dispersion b than in group dispersion d across all orientation situations. It

reinforces the previous findings concerning the MDA discriminating approach. Further, the outcomes summarised in Table 7.3.19 reveal that group dispersion is the major source responsible for interaction effects. The other three levels of group dispersion are not found to be statistically significant when compared with each other. In addition MDA and Logit are very much alike in terms of the impact of indicator's inter-correlation.

### 3. Main effects

All three main effects were significant for the Logit method. The results of the pairwise comparisons and mean misclassification rates are given in Table 7.3.20 and Table 7.3.21. Similarly, the interpretation of the main effects Dist and Disp again shows that outliers and inequality of group dispersion have harmful effects on the classification accuracy.

#### GDR method interaction effects and main effects on training data

##### 1. Three-way interaction effect (Dist by Disp by Orien)

As we have already mentioned before, not all pairwise comparisons are meaningful unless the levels of other factors are fixed. According to previous experience of MDA and Logit methods, the useful comparisons can also be obtained even more clearly by examining the two-way interactions by using graphics. Consequently, we have decided to omit the evaluation of three-way interaction effects hereafter in order to simplify our analysis without impairing it.

##### 2. Two-way interaction effects

###### (1) Dist by Disp effect

For cases with outliers, GDR—the neural network with a nonlinear discriminating feature is also adversely affected. The interaction effect of Dist by Disp displayed by Table 7.3.22 and Figure 7.3.17 shows that outliers always damage the classification ability across all group dispersion situations. Nevertheless, the degree of deterioration does not increase as a result of the inequality of variance-covariance matrices, since the line of normal distribution is almost



parallel to that of symmetric distribution with outliers under these circumstances. Unlike the MDA and Logit methods, when the case is compared to the competitive case (i.e. all other factors hold constant, for example all (a, b) pairs and all (c, d) pairs), GDR shows that a high correlation between attributes improves predictive ability because group dispersion a and c have lower misclassification rates than those in situations b and d irrespective of the distribution of data. A correlation between predictor variables, at least when interacted with the data distribution, is relevant to determine the classification capacity of GDR. This result is consistent with the findings in Chapter Five.

#### (2) Dist by Orien effect

The multiple-comparison results and graphic analysis of the effects of Dist by Orien interaction on GDR method are given in Table 7.3.23 and Figure 7.3.18. Inspecting the results, we find that the source of the interaction occurs between Orientation III and Orientation I as well as between Orientation III and Orientation I across all distributions. There are no significant differences in the misclassification rates for Orientations I and II. Moreover, similarly to MDA, the interaction effect in skewed and normal distributions can be negligible, but in cases with extreme points, the difference between Orientation I and Orientation II becomes larger.

#### (3) Disp by Orien effect

Table 7.3.24 shows the multiple comparison results, and Figure 7.3.19 illustrates the effect of Disp by Orien interaction of GDR method on training data. In terms of within-group correlation factor (group dispersion a vs. b, and group dispersion c vs. d), an interaction is indicated between within-group interrelationships and orientations. With respect to variance-covariance structures (group dispersion a vs. c, and group dispersion b vs. d), interaction effects occur between a and c, but are not strong enough between b and d. As is shown, although equal group dispersion (group b) offers a more efficient solution than unequal group dispersion (group d), it is true only if when the variables' correlation is low. Furthermore, its advantage has been weakened by different orientation schemes (Orientation II and Orientation III), which is quite different from what is displayed in MDA and Logit. Indeed, this outcome gives us a clue that different patterns in the decision

boundary (different orientations between indicators) cause dissimilar impacts on the discrimination between ANNs and statistical methods.

### 3. Main effects

All three main effects were also shown to be significant for the GDR method in the earlier univariate analysis of variance. Mean misclassification rates and results of the pairwise comparisons are given in Table 7.3.25 and Table 7.3.26. For the main effect Dist, the worst classification performances are achieved when the data conforms to the systematic distribution with extreme points. Further, the results for the Disp main effect on the GDR parallel the results of the same main effect on the MDA and Logit. Outliers and inequality of group dispersion between groups are also unfavourable factors for the classification accuracy of the GDR method.

## Proj method interaction effects and main effects on training data

### 2. Two-way interaction effects

#### (1) Dist by Disp effect

Like all the other three alternative techniques, Projection exhibits the worse performance in cases with outliers in the training data. However, the interaction effect of Dist by Disp displayed in Table 7.3.27 and Figure 7.3.20 reveals two new points of interest which are quite different from the previous three. Firstly, there is no strong interaction between data distribution and group dispersion, since the three lines representing different data distributions are nearly parallel over all levels of group dispersion. Secondly, the predictive ability seems to be better for cases with inequality of variance-covariance matrices than with cases of equality of variance-covariance matrices. It can be seen that the group dispersions c and d produce lower misclassification rates compared to their competitive cases a and b for all distribution situations. This surprising outcome demonstrates that Projection works more effectively in a heterogeneous dispersion structure than in a homogeneous one, which contradicts the earlier results obtained in the other three situations.

## (2) Dist by Orien effect

The results of multiple comparisons and the interaction effect are provided in Table 7.3.28 and graphed in Figure 7.3.21. The graph indicates that the source of the interaction exists among Orientation I, II and III across all distributions. This phenomenon is not consistent with the other three alternative methods. The impact of orientation relationships between indicators on the Projection method is different from the impact in the other three discriminators.

## (3) Disp by Orien effect

Table 7.3.29 shows the multiple comparison results, and Figure 7.3.22 illustrates the effects of Disp by Orien on the Proj method. It is also noted that the classification ability is not necessarily better for equal group dispersion than for unequal group dispersion when interacted with the orientation factor. For instance, the performance in group dispersion a (equal group dispersion with high within-group correlation between two attributes) is worse than that in group dispersion c (unequal group dispersion with high within-group correlation between two attributes) across all orientation schemes. This outcome once again supports the findings in the interaction effect of Dist by Disp. This phenomenon in the Projection causes us curiosity to examine whether it also occurs in testing data. Regarding the effect of within-group correlation, there is no clear evidence that a low or high inter-correlation produces better effects in classification ability.

## 3. Main effects

It can be seen from Tables 7.3.30 and 7.3.31, as with the other three methods, that outliers are always the significant factors in causing damage to classification accuracy in the Projection algorithm. However, unlike the other approaches, the data with heterogeneous variance-covariance matrices across groups, as it is unexpectedly presented, has achieved a better performance than data with homogeneous ones. With respect to the orientation factor, although all of these means are identified as statistically significant from each other, we can not interpret directly how this factor influences in predictive performance, since the significant interaction effect is present in every case involved in different orientations.

Table 7.3.1 Misclassification Summary  
in Normal Distribution Data  
Using Training Sample Results

Case	MDA	Logit	GDR	Proj
Type I Error				
Aa1	22.30	23.88	21.00	19.25 ✓
Aa2	18.92	19.58	17.83	15.25 ✓
Aa3	32.08	30.67	35.67	30.08 ✓
Ab1	18.75	18.75	18.25	16.92 ✓
Ab2	24.75	25.50	24.50	24.33 ✓
Ab3	28.42 ✓	29.42	30.25	31.75
Ac1	35.00	34.99	40.17	27.25 ✓
Ac2	31.33	30.24	32.47	24.00 ✓
Ac3	39.58	40.70	46.50	27.17 ✓
Ad1	29.67	28.59	32.58	22.92 ✓
Ad2	35.50	36.17	40.58	30.42 ✓
Ad3	37.58	37.91	45.08	30.58 ✓
Type II Error				
Aa1	21.50 ✓	22.67	22.25	23.25
Aa2	17.33 ✓	18.41	18.00	18.42
Aa3	20.33	24.75	14.25	13.07 ✓
Ab1	18.00	18.08	17.92	17.42 ✓
Ab2	23.67	25.61	23.00	22.83 ✓
Ab3	23.50	26.08	21.17	18.58 ✓
Ac1	19.75	23.50	13.00	9.50 ✓
Ac2	16.17	19.75	10.33	7.58 ✓
Ac3	23.25	27.17	5.33	6.42 ✓
Ad1	14.75	18.75	10.33	8.42 ✓
Ad2	22.17	25.25	13.17	7.75 ✓
Ad3	22.92	26.75	11.08	8.00 ✓
Overall Error				
Aa1	21.92	23.28	21.62	21.25 ✓
Aa2	18.12	19.00	17.92	16.83 ✓
Aa3	26.21	27.71	24.96	21.57 ✓
Ab1	18.38	18.41	18.08	17.17 ✓
Ab2	24.21	25.33	23.75	23.58 ✓
Ab3	25.96	27.75	25.71	25.17 ✓
Ac1	27.37	29.24	26.58	18.37 ✓
Ac2	23.75	25.00	21.40	15.79 ✓
Ac3	31.42	33.94	25.92	16.79 ✓
Ad1	22.21	23.67	21.46	15.67 ✓
Ad2	28.83	30.71	26.87	19.08 ✓
Ad3	30.25	32.33	28.08	19.29 ✓

Table 7.3.2 Misclassification Summary  
in Skewed Distribution Data  
Using Training Sample Results

Case	MDA	Logit	GDR	Proj
Type I Error				
Ba1	28.08 ✓	30.25	34.42	36.00
Ba2	39.67	43.53	33.83	21.83 ✓
Ba3	11.58 ✓	11.58	12.92	12.83
Bb1	29.33 ✓	32.09	35.08	33.83
Bb2	25.67 ✓	27.84	29.58	30.50
Bb3	19.67 ✓	20.76	24.92	26.50
Bc1	38.33	34.83	32.25	31.17 ✓
Bc2	40.42	37.56	36.00	30.67 ✓
Bc3	16.83	13.33	13.92	13.75 ✓
Bd1	37.25	33.17	32.17	29.75 ✓
Bd2	38.17	36.59	33.50	25.83 ✓
Bd3	25.50	22.76	23.50	21.42 ✓
Type II Error				
Ba1	39.83	39.67	29.67	20.50 ✓
Ba2	32.75 ✓	35.92	36.58	35.92
Ba3	7.00	7.99	4.25	2.67 ✓
Bb1	38.33	38.91	29.42	18.92 ✓
Bb2	33.83	34.67	27.67	20.92 ✓
Bb3	22.25	21.17	12.25	7.75 ✓
Bc1	18.00 ✓	22.74	22.33	22.17
Bc2	18.50	24.03	22.17	17.92 ✓
Bc3	4.33	4.65	1.58	1.42 ✓
Bd1	17.92 ✓	23.16	21.25	21.83
Bd2	16.75	20.88	19.58	13.75 ✓
Bd3	9.33	14.83	7.92	4.82 ✓
Overall Error				
Ba1	33.96	34.96	32.04	28.35 ✓
Ba2	36.21	39.72	35.21	28.87 ✓
Ba3	9.29	9.79	8.58	7.75 ✓
Bb1	33.83	35.50	32.25	26.38 ✓
Bb2	29.75	31.25	28.63	25.71 ✓
Bb3	20.96	20.96	18.58	17.13 ✓
Bc1	28.17	28.78	27.29	26.67 ✓
Bc2	29.46	30.80	29.08	24.29 ✓
Bc3	10.58	9.00	7.75	7.58 ✓
Bd1	27.58	28.17	26.71	25.50 ✓
Bd2	27.46	28.73	26.54	19.79 ✓
Bd3	17.42	18.80	15.71	13.12 ✓

Table 7.3.3 Misclassification Summary  
in Symmetric Distribution with Outliers  
Using Training Sample Results

Case	MDA	Logit	GDR	Proj
Type I Error				
Ca1	21.00 ✓	25.76	28.50	29.33
Ca2	34.42	35.32	32.00	29.92 ✓
Ca3	38.83	38.17	37.42 ✓	38.08
Cb1	19.08 ✓	21.49	23.67	25.17
Cb2	37.00	38.90	35.50	32.92 ✓
Cb3	37.58	38.26	35.42	29.42 ✓
Cc1	29.42 ✓	32.39	37.75	39.67
Cc2	38.83	39.41	38.08	34.42 ✓
Cc3	43.00	45.26	45.47	42.92 ✓
Cd1	28.42 ✓	31.83	32.00	40.08
Cd2	39.42 ✓	41.17	39.33	46.50
Cd3	39.08 ✓	41.83	41.08	42.58
Type II Error				
Ca1	37.58	33.42	25.17	20.08 ✓
Ca2	29.92	32.08	29.50	25.75 ✓
Ca3	24.08	27.85	25.08	19.58 ✓
Cb1	34.42	30.42	24.00	16.17 ✓
Cb2	31.33	33.42	32.42	30.25 ✓
Cb3	28.83 ✓	32.59	29.50	30.67
Cc1	38.33	39.50	21.17	8.92 ✓
Cc2	24.08	27.91	24.08	17.50 ✓
Cc3	29.67	34.75	23.50	10.58 ✓
Cd1	37.83	36.09	31.08	12.00 ✓
Cd2	30.50	33.72	27.42	13.42 ✓
Cd3	30.50	32.42	25.92	15.33 ✓
Overall Error				
Ca1	29.29	29.59	26.83	24.71 ✓
Ca2	32.17	33.70	30.75	27.83 ✓
Ca3	31.46	33.01	31.25	28.83 ✓
Cb1	26.75	25.95	23.83	20.67 ✓
Cb2	34.17	36.16	33.96	31.58 ✓
Cb3	33.21	35.42	32.46	30.04 ✓
Cc1	34.12	35.95	29.46	24.29 ✓
Cc2	31.46	33.66	31.08	25.96 ✓
Cc3	36.33	40.00	34.48	26.75 ✓
Cd1	33.12	33.96	31.54	26.04 ✓
Cd2	35.21	37.45	33.37	29.96 ✓
Cd3	34.79	37.13	33.50	28.96 ✓

**Table 7.3.4 The Pattern of Type I and Type II Errors for Each of Four Methods**

Case	MDA	Logit	GDR	Proj
Aa1	-	-	+	+
Aa2	-	-	+	-
Aa3	-	-	-	-
Ab1	-	-	-	-
Ab2	-	-	-	-
Ab3	-	-	-	-
Ac1	-	-	-	-
Ac2	-	-	-	-
Ac3	-	-	-	-
Ad1	-	-	-	-
Ad2	-	-	-	-
Ad3	-	-	-	-
Ba1	-	-	-	-
Ba2	-	-	+	-
Ba3	-	-	-	-
Bb1	+	+	-	-
Bb2	+	+	-	-
Bb3	+	+	-	-
Bc1	-	-	-	-
Bc2	-	-	-	-
Bc3	-	-	-	-
Bd1	-	-	-	-
Bd2	-	-	-	-
Bd3	-	-	-	-
Ca1	+	+	-	-
Ca2	-	-	-	-
Ca3	-	-	-	-
Cb1	+	+	+	-
Cb2	-	-	-	-
Cb3	-	-	-	-
Cc1	+	+	-	-
Cc2	-	-	-	-
Cc3	-	-	-	-
Cd1	+	+	-	-
Cd2	-	-	-	-
Cd3	-	-	-	-

**Table 7.3.1 The Change and Percentage Change in Overall Errors between GDR, Proj and STMs**

Case	GDR		STMs	
	Change in	%Change in	Proj	STMs Method
Aa1	-0.98	-4.34	-1.35	-5.97
Aa2	-0.64	-3.45	-1.73	-9.32
Aa3	-2.00	-7.42	-5.39	-19.99
Ab1	-0.32	-1.71	-1.23	-6.66
Ab2	-1.02	-4.12	-1.19	-4.80
Ab3	-1.15	-4.26	-0.79	-2.94
Ac1	-1.73	-6.09	-9.00	-31.80
Ac2	-2.98	-12.21	-7.96	-32.66
Ac3	-6.76	-20.69	-14.63	-44.77
Ad1	-1.48	-6.45	-6.54	-28.51
Ad2	-2.90	-9.74	-10.69	-35.91
Ad3	-3.21	-6.94	-12.00	-38.35
Ba1	-2.42	-7.02	-6.11	-17.73
Ba2	-2.76	-7.26	-9.10	-23.96
Ba3	-0.96	-10.06	-1.79	-18.76
Bb1	-2.42	-6.97	-8.29	-23.90
Bb2	-1.87	-6.13	-4.79	-15.70
Bb3	-2.38	-11.35	-3.83	-18.27
Bc1	-1.19	-4.16	-1.81	-6.34
Bc2	-1.05	-3.48	-5.84	-19.36
Bc3	-2.04	-20.84	-2.21	-22.57
Bd1	-1.17	-4.18	-2.38	-8.52
Bd2	-1.56	-5.54	-8.31	-29.56
Bd3	-2.40	-13.25	-4.99	-27.55
Ca1	-2.61	-8.87	-4.73	-16.07
Ca2	-2.19	-6.63	-5.11	-15.50
Ca3	-0.99	-3.06	-3.41	-10.56
Cb1	-2.52	-9.56	-5.68	-21.56
Cb2	-1.21	-3.43	-3.59	-10.19
Cb3	-1.86	-5.41	-4.28	-12.46
Cc1	-2.38	-15.91	-10.75	-30.67
Cc2	-1.48	-4.55	-6.60	-20.27
Cc3	-5.69	-9.66	-11.42	-29.91
Cd1	-2.00	-5.96	-7.50	-22.36
Cd2	-2.96	-8.15	-6.37	-17.53
Cd3	2.46	-6.84	-7.00	-19.47

\* denotes the situation in which Type II error is larger than Type I error.  
denotes the situation in which Type II error is lower than Type I error

**Table 7.3.2 The Change and Percentage Change in Type I, Type II Errors between GDR and STMs**

Case	Change in Type I		%Change in Type I		-in Type II		-in Type II	
	Change in	%Change in	Change in	%Change in	Change in	%Change in	Change in	%Change in
Aa1	-2.09	0.17	-9.05	0.75	V			
Aa2	-1.42	0.13	-7.38	0.73	V			
Aa3	4.30	-8.29	13.69	-36.78			V	
Ab1	-0.50	-0.12	-2.67	-0.67	V		V	
Ab2	-0.63	-1.64	-2.49	-6.66	V		V	
Ab3	1.33	-3.62	4.60	-14.60	V		V	
Ac1	5.18	-8.63	14.79	-39.88			V	
Ac2	1.69	-7.63	5.47	-42.48	V		V	
Ac3	6.36	-19.88	15.84	-78.86	V		V	
Ad1	3.45	-6.42	11.84	-38.33	V		V	
Ad2	4.75	-10.54	13.24	-44.45	V		V	
Ad3	7.34	-13.76	19.43	-55.39	V		V	
Ba1	5.26	-10.08	18.02	-25.36	V		V	
Ba2	-7.77	2.25	-18.68	6.54	V		V	
Ba3	3.25	-3.25	11.57	-43.30	V		V	
Bb1	4.37	-9.20	14.23	-23.82	V		V	
Bb2	2.84	-6.58	10.56	-19.21	V		V	
Bb3	4.71	-9.46	23.27	-43.57	V		V	
Bc1	-4.33	1.96	-11.84	9.62	V		V	
Bc2	-2.99	0.91	-7.67	4.26	V		V	
Bc3	-1.16	-2.91	-7.69	-64.81	V		V	
Bd1	-3.04	0.71	-8.63	3.46	V		V	
Bd2	-3.88	0.71	-10.38	3.74	V		V	
Bd3	-0.63	-4.16	-2.61	-34.44	V		V	
Ca1	5.12	-10.33	2.19	-29.10			V	
Ca2	-2.87	-1.50	-8.23	-4.84	V		V	
Ca3	-1.08	-0.89	-2.81	-3.41	V		V	
Cb1	3.39	-8.42	16.68	-25.97	V		V	
Cb2	-2.45	0.05	-6.46	0.14	V		V	
Cb3	-2.50	-1.21	-6.59	-3.94	V		V	
Cc1	6.85	-17.75	22.15	-45.60	V		V	
Cc2	-1.04	-1.92	-2.66	-7.37	V		V	
Cc3	1.34	-8.77	3.04	-27.18	V		V	
Cd1	1.88	-5.88	6.22	-15.91	V		V	
Cd2	-0.97	-4.69	-2.39	-14.61	V		V	
Cd3	0.63	-5.54	1.55	-17.61	V		V	

**Table 7.3.3 The Change and Percentage Change in Type I, Type II Errors between Proj and STMs**

Case	Change in Type I		%Change in Type I		-in Type II		-in Type II	
	Change in	%Change in	Change in	%Change in	Change in	%Change in	Change in	%Change in
Aa1	-3.84	1.17	-16.63	5.28	V			
Aa2	-4.00	0.55	-20.78	3.08	V			
Aa3	-1.30	-9.47	-4.13	-42.01			V	
Ab1	-1.83	-0.62	-9.76	-3.44	V		V	
Ab2	-0.80	-1.81	-3.16	-7.35	V		V	
Ab3	2.83	-6.21	9.79	-25.05	V		V	
Ac1	-7.75	-12.13	-22.13	-56.07	V		V	
Ac2	-6.79	-10.38	-22.04	-57.8	V		V	
Ac3	-12.97	-18.79	-32.31	-74.53	V		V	
Ad1	-6.21	-8.33	-21.32	-49.73	V		V	
Ad2	-5.42	-15.96	-15.11	-67.31	V		V	
Ad3	-7.17	-16.84	-18.98	-67.79	V		V	
Ba1	6.84	-19.25	23.44	-48.43	V		V	
Ba2	-19.77	1.59	-47.52	4.62	V		V	
Ba3	1.25	-4.83	10.79	-4.38	V		V	
Bb1	3.12	-19.7	10.60	-51.01	V		V	
Bb2	3.75	-13.33	14.00	-38.92	V		V	
Bb3	6.29	-13.96	31.09	-64.30	V		V	
Bc1	-5.41	1.80	-14.79	8.84	V		V	
Bc2	-8.32	-3.35	-21.34	-15.73	V		V	
Bc3	-1.33	-3.07	-8.82	-68.37	V		V	
Bd1	-5.46	1.29	-15.51	6.28	V		V	
Bd2	-11.58	-5.73	-30.98	-30.33	V		V	
Bd3	-2.71	-7.26	-11.23	-60.10	V		V	
Ca1	5.95	-15.42	25.45	-43.44	V		V	
Ca2	-4.95	-5.25	-14.20	-16.94	V		V	
Ca3	-0.42	-6.39	-1.09	-24.59	V		V	
Cb1	4.89	-16.25	24.08	-50.12	V		V	
Cb2	-5.03	-2.22	-13.25	-6.56	V		V	
Cb3	-8.50	-0.04	-22.42	-0.13	V		V	
Cc1	10.25	-30.00	33.17	-77.08	V		V	
Cc2	-4.70	-8.49	-12.01	-32.68	V		V	
Cc3	-1.21	-21.69	-2.74	-67.21	V		V	
Cd1	9.96	-24.96	33.05	-67.53	V		V	
Cd2	6.21	-18.69	15.40	-58.21	V		V	
Cd3	2.13	-16.13	5.25	-51.27	V		V	

Table 7.3.5 Multivariate Analysis of Variance of MDA, Logit, GDR and Proj Methods on Overall Error Rates Using Training Data

	Wilks'	Hypoth. df	Error of	Multivariate
Dist	0.46986056	8	1362.000	78.1220*
Disp	0.72698127	12	1802.048	19.2336*
Orien	0.77178004	8	1362.000	23.5440*
Dist by Disp	0.64277772	24	2376.940	13.3763*
Dist by Orien	0.34139947	16	2081.127	54.8369*
Disp by Orien	0.77929476	24	2376.935	7.33850*
Dist by Disp by Orien	0.68956694	48	2625.322	5.54050*

Table 7.3.6 Univariate Analysis of Variance for MDA Using Training Samples

Source of Variance	Sum of Square	DF	Mean Square of	F Value
Dist	9117.548	2	4558.774	258.03*
Disp	413.8	3	137.93	7.81*
Orien	1595.008	2	797.5	45.14*
Dist by Disp	2618.175	6	436.36	24.70*
Dist by Orien	14496.59	4	3624.147	205.13*
Disp by Orien	1102.024	6	183.67	10.40*
Dist by Disp by	2904.193	12	242.02	13.70*

Table 7.3.7 Univariate Analysis of Variance for Logit Using Training Samples

Source of Variance	Sum of Square	DF	Mean Square of	F Value
Dist	10145.63	2	5072.819	246.73*
Disp	557.06	3	185.69	9.03*
Orien	1740.661	2	870.33	42.33*
Dist by Disp	3814.801	6	635.8	30.92*
Dist by Orien	17982.49	4	4495.624	218.65*
Disp by Orien	1477.452	6	246.24	11.98*
Dist by Disp by	3834.573	12	319.55	15.54*

Table 7.3.8 Univariate Analysis of Variance for GDR Using Training Samples

Source of Variance	Sum of Square	DF	Mean Square of	F Value
Dist	8470.796	2	4235.398	267.45*
Disp	261.222	3	87.07	5.50*
Orien	2244.419	2	1122.209	70.86*
Dist by Disp	1679.711	6	279.95	17.68*
Dist by Orien	15321.89	4	3830.472	241.88*
Disp by Orien	930.632	6	155.11	9.79*
Dist by Disp by	2839.560	12	236.63	14.94*

Table 7.3.9 Univariate Analysis of Variance for Proj Using Training Samples

Source of Variance	Sum of Square	DF	Mean Square of	F Value
Dist	8342.654	2	4171.327	301.58*
Disp	1142.481	3	380.83	27.53*
Orien	1873.363	2	936.68	67.72*
Dist by Disp	10703.83	6	82.69	5.98*
Dist by Orien	10703.83	4	2675.957	193.47*
Disp by Orien	1612.779	6	268.7966	19.43*
Dist by Disp by	1558.982	12	129.92	9.39*

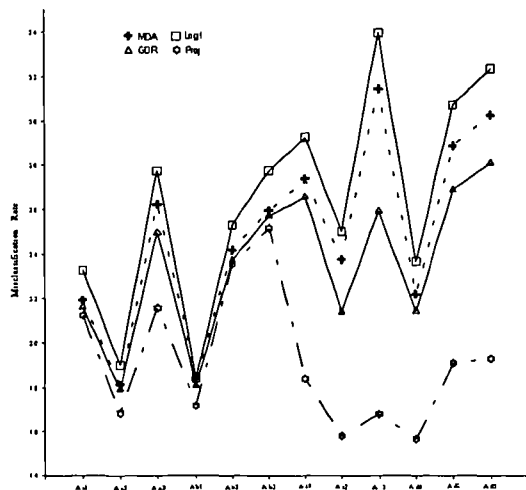


Figure 7.3.1 Comparison on Normal Distribution for Training Data

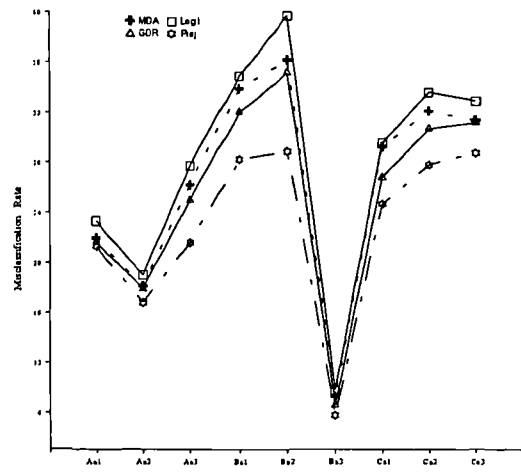


Figure 7.3.4 Comparison on Group Dispersion (a) for Training Data

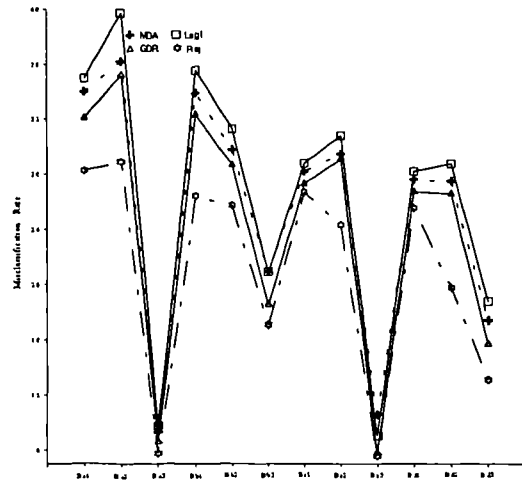


Figure 7.3.2 Comparison on Skewed Distribution for Training Data

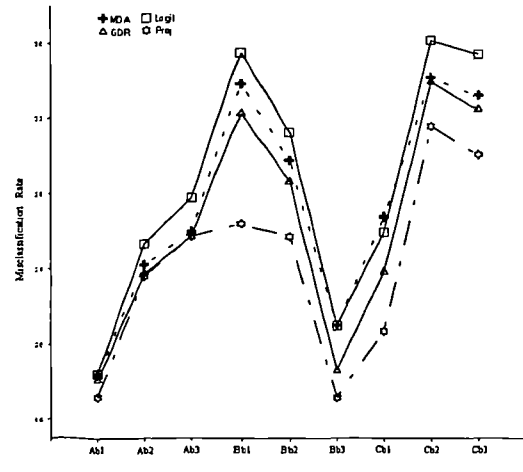


Figure 7.3.5 Comparison on Group Dispersion (b) for Training Data

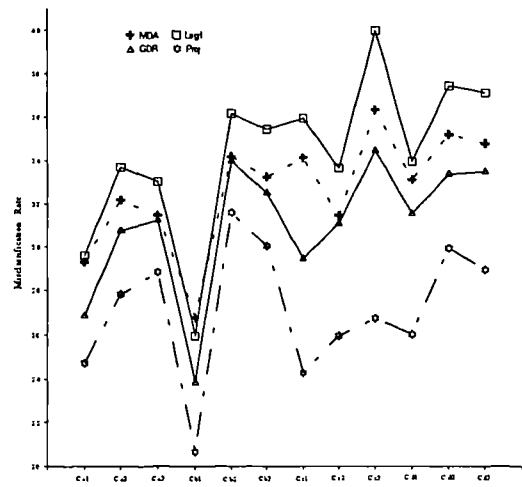


Figure 7.3.3 Comparison on Symmetric Distribution with Outliers for Training Data

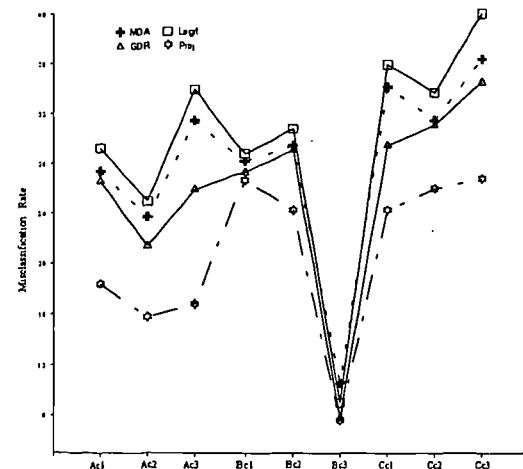


Figure 7.3.6 Comparison on Group Dispersion (c) for Training Data

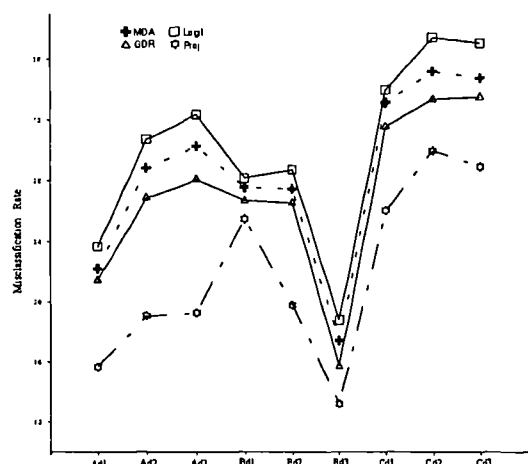


Figure 7.3.7 Comparison on Group Dispersion (d) for Training Data

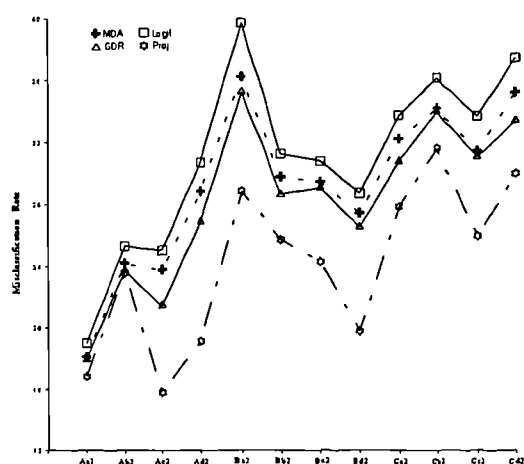


Figure 7.3.9 Comparison On Orientation Scheme II for Training Data

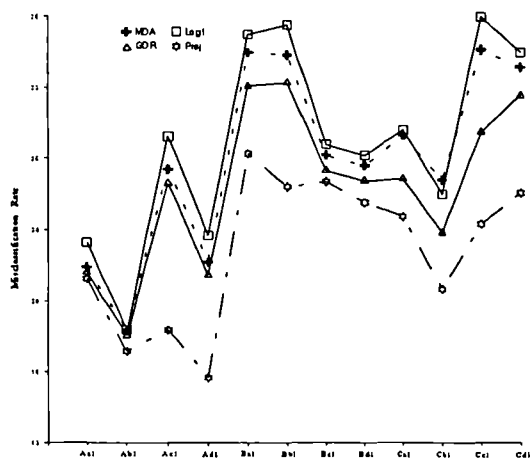


Figure 7.3.8 Comparison on Orientation Scheme I for Training Data

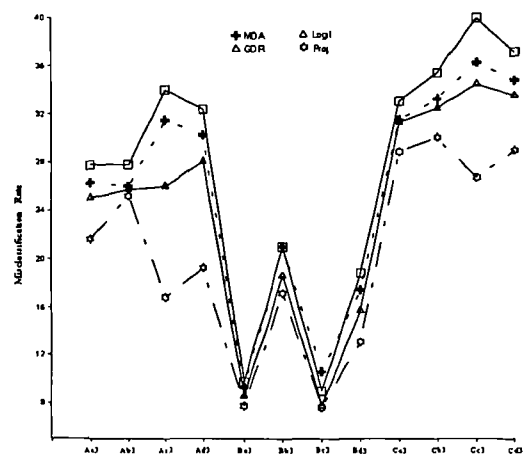


Figure 7.3.10 Comparison on Orientation Scheme III for Training Data



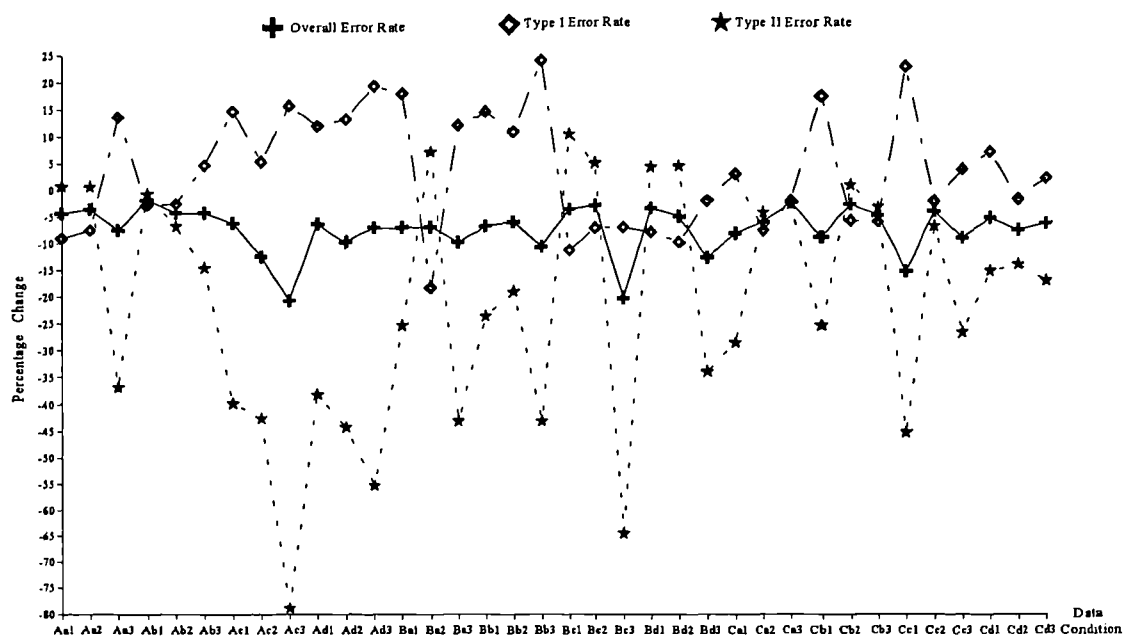


Figure 7.3.I The Plot of Percentage Change in Overall, Type I and Type II Error Rates between GDR and Statistical Methods Using Training Samples

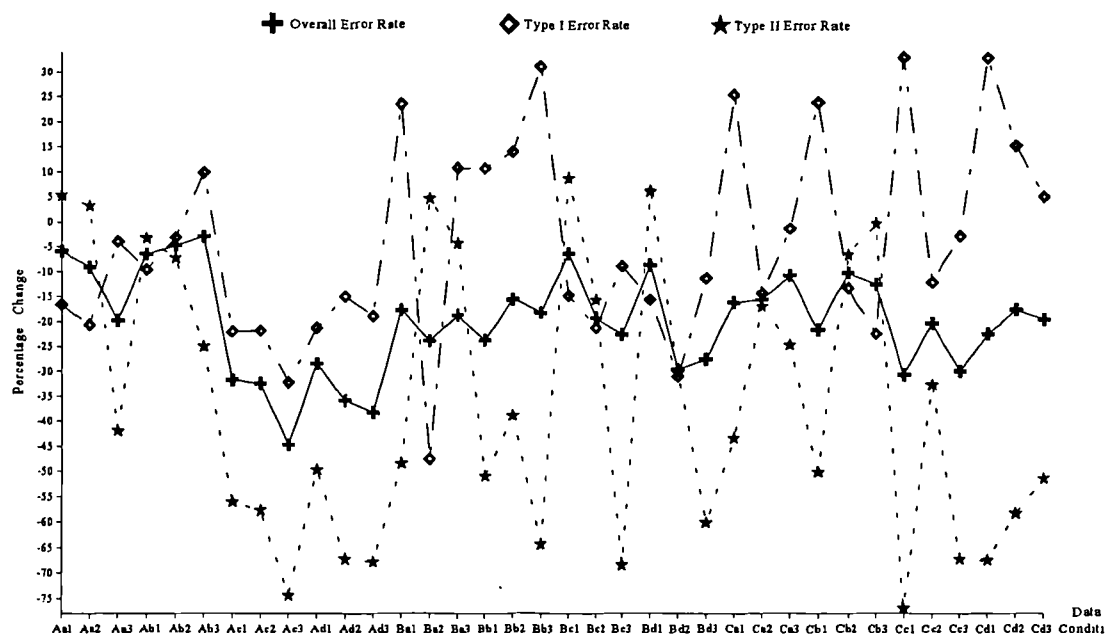


Figure 7.3.II The Plot of Percentage Change in Overall, Type I and Type II Error Rates between Proj and Statistical Methods Using Training Samples

Table 7.3.10 The Three-Way Interaction Effects Dist by Disp by Orien for MDA Method Using Training Samples

	Aa1	Aa2	Aa3	Ab1	Ab2	Ab3	Ac1	Ac2	Ac3	Ad1	Ad2	Ad3	Ba1	Ba2	Ba3	Bb1	Bb2	Bb3	Bc1	Bc2	Bc3	Bd1	Bd2	Bd3	Ca1	Ca2	Ca3	Cb1	Cb2	Cb3	Cc1	Cc2	Cc3	Cd1	Cd2	Cd3	
Aa1	—																																				
Aa2		—																																			
Aa3			—																																		
Ab1				—																																	
Ab2					—																																
Ab3						—																															
Ac1							—																														
Ac2								—																													
Ac3									—																												
Ad1										—																											
Ad2											—																										
Ad3												—																									
Ba1													—																								
Ba2														—																							
Ba3															—																						
Bb1																—																					
Bb2																	—																				
Bb3																		—																			
Bc1																			—																		
Bc2																				—																	
Bc3																					—																
Bd1																						—															
Bd2																							—														
Bd3																								—													
Ca1																									—												
Ca2																										—											
Ca3																											—										
Cb1																												—									
Cb2																													—								
Cb3																														—							
Cc1																															—						
Cc2																																—					
Cc3																																—					
Cd1																																	—				
Cd2																																		—			
Cd3																																			—		

\* denotes significant using 0.05 family-wise alpha

172

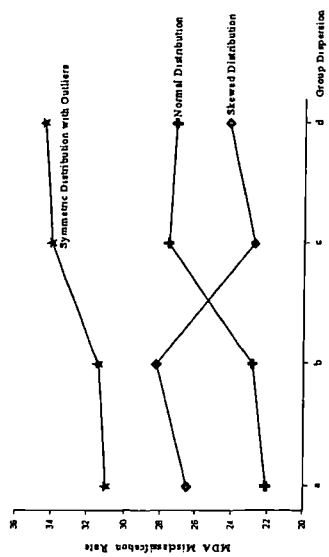


Figure 7.3.11 Interaction Effects of Dist by Disp on MDA Using Training samples

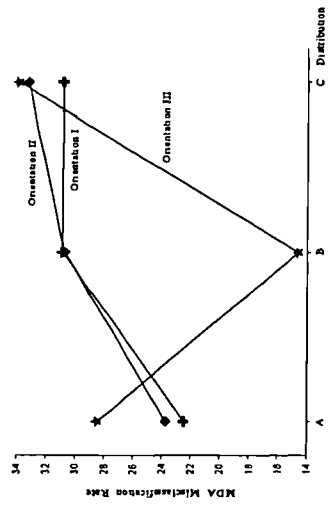


Figure 7.3.12 Interaction Effects Of Dist by Orien on MDA Using Training Samples

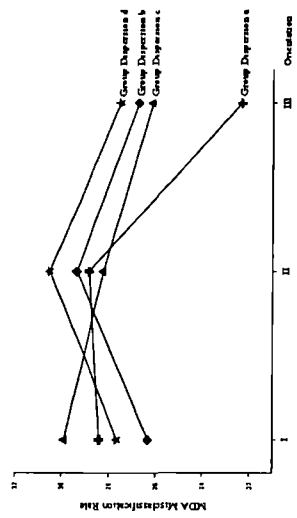


Figure 7.3.13 Interaction Effects of Disp by Orien on MDA Using Training Samples

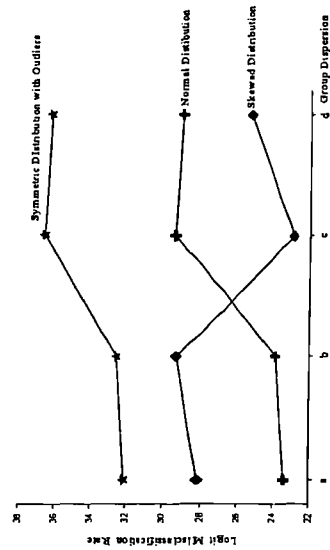


Figure 7.3.14 Interaction Effects of Dist by Disp on Logit Using Training Samples

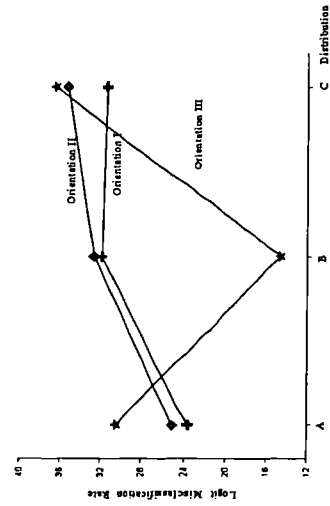


Figure 7.3.15 Interaction Effects of Dist by Disp on Logit Using Training Samples

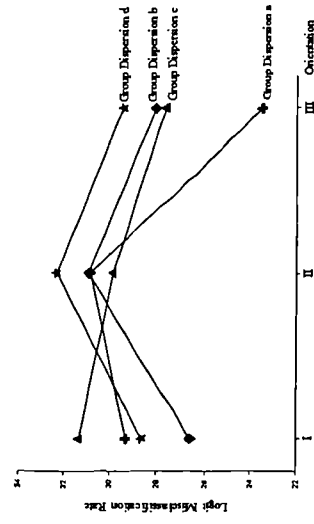


Figure 7.3.16 Interaction Effects of Disp by Orien on Logit Using Training Samples

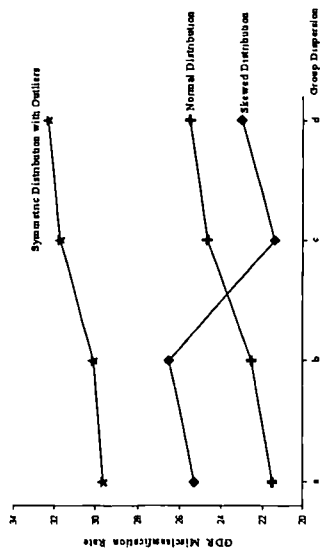


Figure 7.3.17 Interaction Effects of Dist by Disp on GDR Using Training Samples

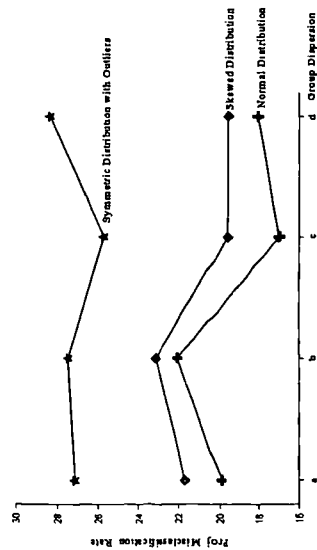


Figure 7.3.20 Interaction Effects of Dist by Disp on Proj Using Training Samples

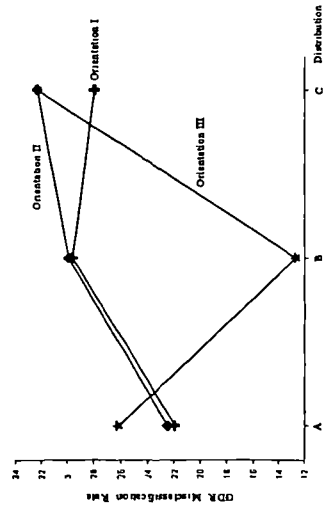


Figure 7.3.18 Interaction Effects of Dist by Orien on GDR Using Training Samples

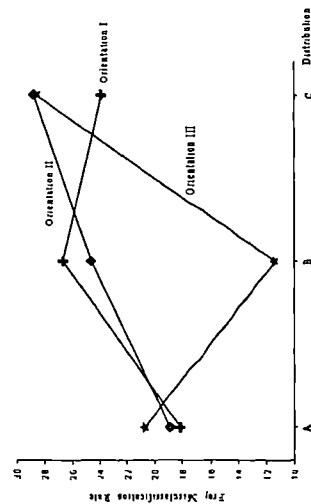


Figure 7.3.21 Interaction Effects of Dist by Orien on Proj Using Training Samples

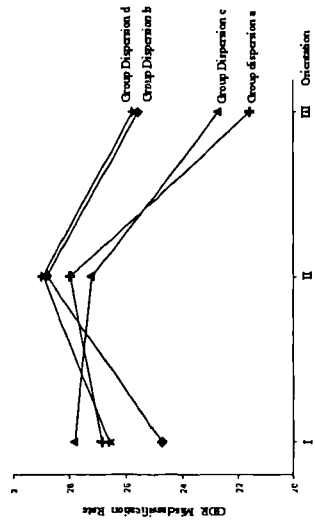


Figure 7.3.19 Interaction Effects of Disp by Orien on GDR Using Training Samples

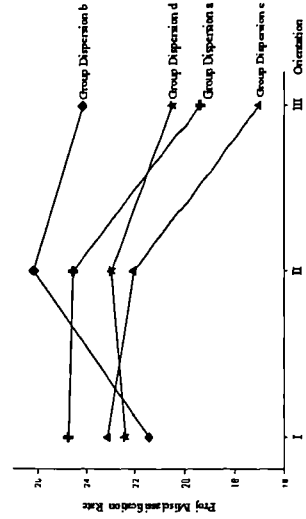


Figure 7.3.22 Interaction Effects of Disp by Orien on Proj Using Training Samples

Table 7.3.11 Multiple Comparison on MDA  
Dist by Disp Effect for Training Data

Aa	Ab	Ac	Ad	Ba	Bb	Bc	Bd	Ca	Cb	Cc	Cd
Aa	—	*	*	*	*	*	*	*	*	*	*
Ab	—	*	*	*	*	*	*	*	*	*	*
Ac	—	*	*	*	*	*	*	*	*	*	*
Ad	—	*	*	*	*	*	*	*	*	*	*
Ba	—	*	*	*	*	*	*	*	*	*	*
Bb	—	*	*	*	*	*	*	*	*	*	*
Bc	—	*	*	*	*	*	*	*	*	*	*
Bd	—	*	*	*	*	*	*	*	*	*	*
Ca	—	*	*	*	*	*	*	*	*	*	*
Cb	—	*	*	*	*	*	*	*	*	*	*
Cc	—	*	*	*	*	*	*	*	*	*	*
Cd	—	*	*	*	*	*	*	*	*	*	*

Table 7.3.13 Multiple Comparison on MDA  
Dist by Orien Effect for Training Data

a1	a2	a3	b1	b2	b3	c1	c2	c3	d1	d2	d3
a1	—	*	*	*	*	*	*	*	*	*	*
a2	—	*	*	*	*	*	*	*	*	*	*
a3	—	*	*	*	*	*	*	*	*	*	*
b1	—	*	*	*	*	*	*	*	*	*	*
b2	—	*	*	*	*	*	*	*	*	*	*
b3	—	*	*	*	*	*	*	*	*	*	*
c1	—	*	*	*	*	*	*	*	*	*	*
c2	—	*	*	*	*	*	*	*	*	*	*
c3	—	*	*	*	*	*	*	*	*	*	*
d1	—	*	*	*	*	*	*	*	*	*	*
d2	—	*	*	*	*	*	*	*	*	*	*
d3	—	*	*	*	*	*	*	*	*	*	*

Table 7.3.12 Multiple Comparison  
on MDA Dist by Orien Effect  
for Training Data

A1	A2	A3	B1	B2	B3	C1	C2	C3
A1	—	*	*	*	*	*	*	*
A2	—	*	*	*	*	*	*	*
A3	—	*	*	*	*	*	*	*
B1	—	*	*	*	*	*	*	*
B2	—	*	*	*	*	*	*	*
B3	—	*	*	*	*	*	*	*
C1	—	*	*	*	*	*	*	*
C2	—	*	*	*	*	*	*	*
C3	—	*	*	*	*	*	*	*

Table 7.3.17 Multiple Comparison on Logit  
Dist by Disp Effect for Training Data

Aa	Ab	Ac	Ad	Ba	Bb	Bc	Bd	Ca	Cb	Cc	Cd
Aa	—	*	*	*	*	*	*	*	*	*	*
Ab	—	*	*	*	*	*	*	*	*	*	*
Ac	—	*	*	*	*	*	*	*	*	*	*
Ad	—	*	*	*	*	*	*	*	*	*	*
Ba	—	*	*	*	*	*	*	*	*	*	*
Bb	—	*	*	*	*	*	*	*	*	*	*
Bc	—	*	*	*	*	*	*	*	*	*	*
Bd	—	*	*	*	*	*	*	*	*	*	*
Ca	—	*	*	*	*	*	*	*	*	*	*
Cb	—	*	*	*	*	*	*	*	*	*	*
Cc	—	*	*	*	*	*	*	*	*	*	*
Cd	—	*	*	*	*	*	*	*	*	*	*

Table 7.3.19 Multiple Comparison on Logit  
Dist by Orien Effect for Training Data

a1	a2	a3	b1	b2	b3	c1	c2	c3	d1	d2	d3
a1	—	*	*	*	*	*	*	*	*	*	*
a2	—	*	*	*	*	*	*	*	*	*	*
a3	—	*	*	*	*	*	*	*	*	*	*
b1	—	*	*	*	*	*	*	*	*	*	*
b2	—	*	*	*	*	*	*	*	*	*	*
b3	—	*	*	*	*	*	*	*	*	*	*
c1	—	*	*	*	*	*	*	*	*	*	*
c2	—	*	*	*	*	*	*	*	*	*	*
c3	—	*	*	*	*	*	*	*	*	*	*
d1	—	*	*	*	*	*	*	*	*	*	*
d2	—	*	*	*	*	*	*	*	*	*	*
d3	—	*	*	*	*	*	*	*	*	*	*

Table 7.3.18 Multiple Comparison  
on Logit Dist by Orien Effect  
for Training Data

A1	A2	A3	B1	B2	B3	C1	C2	C3
A1	—	*	*	*	*	*	*	*
A2	—	*	*	*	*	*	*	*
A3	—	*	*	*	*	*	*	*
B1	—	*	*	*	*	*	*	*
B2	—	*	*	*	*	*	*	*
B3	—	*	*	*	*	*	*	*
C1	—	*	*	*	*	*	*	*
C2	—	*	*	*	*	*	*	*
C3	—	*	*	*	*	*	*	*

Table 7.3.14 Main Effects on MDA for Training Data

A	B	C	a	b	c	d	I	II	III
A	—	*	a	—	*	*	I	—	*
B	—	*	b	—	—	—	II	—	*
C	—	—	c	—	—	—	III	—	—
			d	—	—	—	—	—	—

Table 7.3.15 Mean Error on MDA for Training Data

Le	Mean Error Rate	Le	Mean Error Rate	Leve	Mean Error Rate
A	26.36	a	27.86	I	28.96
B	26.37	b	28.53	II	30.96
C	34.34	c	29.53	III	27.15
		d	30.11		

Table 7.3.20 Main Effects on Logit for Training Data

A	B	C	a	b	c	d	I	II	III
A	—	*	a	—	*	*	I	—	*
B	—	*	b	—	*	*	II	—	*
C	—	—	c	—	—	—	III	—	—
			d	—	—	—	—	—	—

Table 7.3.21 Mean Error on Logit for Training Data

Le	Mean Error Rate	Le	Mean Error Rate	Leve	Mean Error Rate
A	26.364	a	27.861	I	28.956
B	26.371	b	28.527	II	30.959
C	34.337	c	29.526	III	27.152
		d	30.105		

\* denotes significant using a 0.05 family-wise alpha

Table 7.3.22 Multiple Comparison on GDR  
Dist by Disp Effect for Training Data

Aa	Ab	Ac	Ad	Ba	Bb	Bc	Bd	Ca	Cb	Cc	Cd
Aa	—			*				*	*	*	*
Ab	—							*	*	*	*
Ac		—						*	*	*	*
Ad			—		*			*	*	*	*
Ba				—				*	*	*	*
Bb					—	*		*	*	*	*
Bc						—		*	*	*	*
Bd							—	*	*	*	*
Ca								—			
Cb									—		
Cc										—	
Cd											—

Table 7.3.23 Multiple Comparison on GDR  
Dist by Orient Effect for Training Data

A1	A2	A3	B1	B2	B3	C1	C2	C3
A1	—	*	*	*	*	*	*	*
A2	—	*	*	*	*	*	*	*
A3		—	*	*	*	*	*	*
B1			—	*	*	*	*	*
B2				—	*	*	*	*
B3					—	*	*	*
C1						—	*	*
C2							—	*
C3								—

Table 7.3.24 Multiple Comparison on GDR  
Dist by Orient Effect for Training Data

a1	a2	a3	b1	b2	b3	c1	c2	c3	d1	d2	d3
a1	—	*									
a2	—	*									
a3		—	*								
b1			—	*							
b2				—	*						
b3					—	*					
c1						—	*				
c2							—	*			
c3								—	*		
d1									—	*	
d2										—	*
d3											—

Table 7.3.25 Main Effects on GDR or Training Data

A	B	C	a	b	c	d	I	II	III
A	—	*	a	—	—	—	I	—	*
B	—	*	b	—	—	—	II	—	*
C	—	—	c	—	—	—	III	—	—
			d	—	—	—	—	—	—

Table 7.3.26 Mean Error on GDR for Training Data

Le	Mean Error Rate	Le	Mean Error Rate	Leve	Mean Error Rate
A	23.530	a	25.463	I	26.476
B	24.034	b	26.361	II	28.214
C	31.044	c	25.895	III	23.915
		d	27.088		

Table 7.3.27 Multiple Comparison on Proj  
Dist by Disp Effect for Training Data

Aa	Ab	Ac	Ad	Ba	Bb	Bc	Bd	Ca	Cb	Cc	Cd
Aa	—			*				*	*	*	*
Ab	—	*		*				*	*	*	*
Ac		—	*	*				*	*	*	*
Ad			—	*	*			*	*	*	*
Ba				—				*	*	*	*
Bb					—	*		*	*	*	*
Bc						—	*	*	*	*	*
Bd							—	*	*	*	*
Ca								—			
Cb									—		
Cc										—	
Cd											—

Table 7.3.28 Multiple Comparison  
on Proj Dist by Orient Effect  
for Training Data

A1	A2	A3	B1	B2	B3	C1	C2	C3
A1	—	*	*	*	*	*	*	*
A2	—	*	*	*	*	*	*	*
A3		—	*	*	*	*	*	*
B1			—	*	*	*	*	*
B2				—	*	*	*	*
B3					—	*	*	*
C1						—	*	*
C2							—	*
C3								—

Table 7.3.29 Multiple Comparison on Proj  
Dist by Orient Effect for Training Data

a1	a2	a3	b1	b2	b3	c1	c2	c3	d1	d2	d3
a1	—	*				*					*
a2	—	*				*					*
a3		—	*			*					*
b1			—	*		*					*
b2				—	*	*					*
b3					—	*					*
c1						—	*				*
c2							—	*			*
c3								—	*		*
d1									—	*	*
d2										—	*
d3											—

Table 7.3.30 Main Effects on Proj for Training Data

A	B	C	a	b	c	d	I	II	III
A	—	*	a	—	—	—	I	—	*
B	—	*	b	—	—	—	II	—	*
C	—	—	c	—	—	—	III	—	—
			d	—	—	—	—	—	—

Table 7.3.31 Mean Error on Proj for Training Data

Le	Mean Error Rate	Le	Mean Error Rate	Leve	Mean Error Rate
A	19.214	a	22.878	I	22.913
B	20.919	b	24.157	II	24.107
C	27.135	c	20.722	III	20.249
		d	21.934		

\* denotes significant using a 0.05 family-wise alpha

## 7.4 The Results and Analyses of Testing Samples

The data for the testing samples was gathered by using the four techniques developed for each set of data conditions to classify 120 new observations (60 from each group) conforming to the same data conditions. The primary objective of the discriminating process is to classify new observations into groups based on what is known about pre-existing members of those groups. Therefore, the results of the analyses of testing samples are the most important.

### 7.4.1 Descriptive Statistics and Graphic Analyses

The mean misclassification rates and standard deviations of 20 replications for each of the four discriminating methods under various combinations of data conditions are given in Tables 7.4.1 to 7.4.3.

First, we are interested in the pattern of Type I and Type II error provided by different methods. Similarly, the sign of + means that the Type II error is larger than the Type I error and the sign of – denotes the opposite situation. The sign of \* then describes the tie position. Table 7.4.4 again indicates that in almost all cases MDA and Logit generate the same signs over testing conditions except for the cases Ab2 and Ab3. In effect, the differences between Type I and Type II errors in these two cases are very small and can be negligible. On the other hand, GDR and Projection always have the same pattern between them. However, a further inspection finds that opposite signs between STMs and ANNs occur frequently in the data condition having skewed distribution with equal variance-covariance matrices and in the presence of outlier data.

The comparisons of Overall error rates among the four methods are then discussed. The differences in numbers and in percentages between GDR and statistical methods, as well as between Proj and statistical methods, are calculated and reported in Table 7.4.I. Similarly to the training samples, most of the significant differences (more than or close to ten) occur in situations with Orientation III involved (for instance, cases Ac3, Bb3, and Bc3) on the GDR. As for the Projection, in addition to the fact that Orientation III still plays a crucial role in determining differences in classification ability, the impacts of unequal group dispersions and outliers are evident as well. These results are demonstrated in cases Ac1, Ac2, Ac3, Ad1, Ad2, Ad3, Cc1, Cc2 Cc3 and Cd1.



The descriptive statistics of Table 7.4.1 to 7.4.3 are also graphed in Figure 7.4.1 to 7.4.10 in terms of different data distributions, group dispersion levels and orientation schemes.

As Figure 7.4.1 shows, under normal distribution, when the variance-covariance structures across two groups are identical (cases Aa1, Aa2, Aa3, Ab1, Ab2 and Ab3), there is no clear evidence of which method has the best performance. However, it is conspicuous that STMs present their successful generalisation capacities in contrast to ANNs. Particularly, the Projection algorithm, though it has exhibited its remarkable superiority in learning ability, demonstrates the worst level of generalisation. At the same time, if group dispersion are heterogeneous, Projection yields the significantly lowest misclassification rates, while the two ANNs are proved to clearly outperform the two STMs under such circumstances.

For skewed distribution, whatever the covariance-variance matrices or the orientation schemes, no discriminating method dominates the other three over all conditions. With respect to the data with outliers, it is observed that ANNs undoubtedly perform better than statistical methods, especially as the difference of group dispersion increases (cases Cc1, Cc2, Cc3, Cd1, Cd2, and Cd3). For the remaining cases, except for Cb2, notwithstanding the fact that MDA or Logit still yield higher misclassification rates than GDR or Proj, this disadvantages become smaller. This implies that there exist interaction effects between factors affecting predictive performances.

Figures 7.4.4 to 7.4.7 are provided for assessing the effect of group dispersion of predictor variables. Figures 7.4.4 and 7.4.5 represent the data having equal variance-covariance matrices with either a high or low within-group correlation between two attributes respectively. Figures 7.4.6 and 7.4.7 describe the situations of unequal variance-covariance matrices with high or low correlation. In terms of different within-group correlations, there is no conclusive evidence that the classification accuracy will be influenced by the high or low inter-relationship of indicators (Figure 7.4.5 vs. 7.4.6, Figure 7.4.7. vs. 7.4.8). As for the group dispersion structures, in some cases, the equality or inequality of variance-covariance matrices appears to be a factor causing the relative effectiveness of ANNs and statistical methods. However, this impact seems inconsistent under certain conditions for different approaches, and thus needs to be further explored by multivariate and univariate analysis.

Considering the angle of different orientation schemes, we are not seeking to interpret them directly through graphics since some of them are not competitive cases. Nonetheless, examining Table 7.4.I, we can see that ANNs obviously gain an advantage over STMs from the Orientation III scheme. This phenomenon is always related to the ability of multi-layered Perceptron (MLP) to cope with nonlinear boundaries.

Comparing the testing results with the training results, we find that the overwhelming dominance of the Projection approach in training data does not exist in the same way as in testing data. Nor does the ANNs over STMs. Under some circumstances (especially, in cases Aa1, Aa2, Ab1, Ab2, Ab3 and Bd1), ANNs yield even worse results than STMs in testing samples. We shall bear in mind that the outperformances of ANNs should never obscure the potential risk of their overfitting tendency.

#### **7.4.2 Change in Type I and Type II Error Rates**

Table 7.4.II and Table 7.4.III report the value and percentage change of Type I and Type II error rates relative to Overall error rates for GDR and Proj individually. The results of change in percentages are also illustrated by Figures 7.4.I and 7.4.II.

In GDR the improvement in overall accuracy is still shown in favour of classifying nonbankrupt firms. 26 out of 36 cases exhibit reduced Type II errors when Overall performance is improved, while only 10 cases indicate a decrease in Type I errors. Surprisingly, none of them indicates a decrease in both Type I and Type II error rates. In Projection, 24 of 36 cases display an improvement in Type II error accompanied by an improvement in Overall accuracy. 17 cases exhibit a decrease in Type I error. 7 cases have reductions in both Type I and Type II errors, but 2 cases have growth in both errors as a result of worse overall performances.

Further, when data has a normal distribution and homogeneous dispersion (Aa1- Ab3), MDA again turns out to be the most effective method for achieving better generalisation. It also yields lower Type I errors than GDR under a normal model.

Combining the results of training and testing samples, ANN methods have proved their superiority under comprehensive data conditions. However, the improvement in overall performances does not necessarily benefit the Type I error as well. Thus, we cannot unconditionally conclude that ANNs outperform STMs, since Type I error cost is generally believed to be much higher than Type II error cost in the real world.

### 7.4.3 Multivariate Analysis

The MANOVA model seeks to determine the main effects as well as interaction effects among the four methods. The results of the MANOVA for the testing samples data are similar to the results of the MANOVA for the training sample data. All three main effects (Dist, Disp, Orien), two-way interaction effects (Dist by Disp, Dist by Orien and Disp by Orien) and one three-way interaction effect (Dist by Disp by Orien) are significant at  $p < 0.001$ . These results are reported in Table 7.4.5. The results support the hypotheses for the testing samples as they did for the training samples. The source of the significance and interactions is also traced by examining the univariate analysis.

### 7.4.4 Univariate Analysis

The significant main effects and interactions found in the MANOVA are examined in the univariate analysis for each of the four dependent variables different methods. A per comparison alpha of 0.0125 was found to control Type I error. The results are displayed in Tables 7.4.6 to 7.4.9. Examination of these analyses again indicates significance for almost all main and interaction effects for each of four methods except for Disp main effect in GDR. The interpretation of the interaction effects will allow the factors involved in both main effects and interactions to be clearly understood. Therefore the levels at which the main effects and interaction effects are significant must be found individually. This will be accomplished by multiple comparisons using the Tukey procedure. The interaction effects are examined first.

### 7.4.5 Interaction Effects and Main Effects

MDA interaction effects and main effects on testing data

#### 1. Two-way interaction effects

##### (1) Dist by Disp effect

The interaction effects of Dist by Disp are graphed in Figure 7.4.11. The better classification was obviously attained in normal distributions compared to outlier cases. In Table 7.4.10 multiple comparisons of the means of the effects are made using the Tukey procedure and a 0.05 familywise alpha. This table reports the

source of the interaction for the MDA to be between skewed distribution and normal distribution, as well as skewed distribution and distribution having outliers across all group dispersions. It is outliers, not skew data, that create a difficulty in classifying on both training and testing samples for the MDA method.

(2) Dist by Orien effect

The multiple comparisons and interaction effect are provided in Table 7.4.11 and graphed in Figure 7.4.12 respectively. Unlike the results in the training data, it is found that orientation is a relevant factor when mixed with different data distributions. Thus the corresponding main effects have little practical meaning.

(3) Disp by Orien effect

Table 7.4.12 displays multiple-comparison results and Figure 7.4.13 illustrates the effects of Disp by Orien interaction on the MDA method. The graph indicates that the misclassification rates are lower for group dispersion b than those for group dispersion d over all ranges of orientations. This result again clearly shows that when the group dispersion is heterogeneous and the predictor variables have a low inter-correlation, MDA does not perform as well as with homogenous data. The same evidence in both training and testing data show that in the MDA procedure, the equality of variance-covariance matrices across groups is indeed the necessary factor which can improve classification accuracy. On the other hand, comparing group dispersion a and c, we find that highly inter-correlated predictor variables also have certain impact on predictive ability.

### 3. Main effects

In the light of the testing results in MDA, the main effects are reported in Tables 7.4.13 and 7.4.14. For data distribution and the orientation scheme, the results conform with the effects involved in interactions. However, it is noted that there are no significant differences among various levels of group dispersion. This outcome contradicts the previous findings. This issue arises because the differences between group dispersions are cancelled out by the interaction effects with data distribution and the orientation between two attributes. Put another way, a significant interaction obscures the significance of the main effect. This is the reason why, before we analyse the main effects, the significant interaction effects should be first examined.

## Logit method interaction effects and main effects on testing data

### 1. Two-way interaction effects

#### (1) Dist by Disp effect

Table 7.4.15 and Figure 7.4.14 present the multiple comparisons and interaction effects of Dist by Disp in testing data for Logit method. Through graphic analysis, outliers are also demonstrated to be detrimental to classification power regardless of the group dispersion of the data. Further, the multiple comparisons shows that the main source of the interaction is traced to differences in mean misclassification rates between skewed distribution and normal distribution as well as skewed distribution and data having extreme points. There is no adequate evidence that the Logit method work less effectively in skewed data.

#### (2) Dist by Orien effect

The testing sample results of multiple comparisons and graphic analysis of the effects of Dist by Orien interaction on Logit method are given in Table 7.4.16 and Figure 7.4.14. Inspection of the results shows that, in addition to the fact that interactions still occur among all orientations, the interaction between Orientation I and Orientation II has some clear implications. The difference between them is much smaller in normal and skewed cases than in cases where the data consists of extreme points. That is, the classification performance is affected when the orientation between two attributes is interacted with different data distributions.

#### (3) Disp by Orien effect

The graph of the interaction effects of Disp by Orien using testing data on the Logit method is shown in Figure 7.4.16. The significance of average misclassification rates in paired comparison is shown in Table 7.4.17 for the testing data as it was for the training data. Very similarly to the results in MDA, the misclassification rates are lower for the homogeneous groups with low inter-correlated independent variables (group dispersion b vs. group dispersion d). These findings again prove that the Logit formulation classifies homogeneous groups better than it does heterogeneous groups, and that even the equality of variance-covariance matrices is not a necessary assumption to the development of Logit procedure.

### 3. Main effects

The main effects of the Logit method on testing data have all been previously shown as significant by examining univariate analyses. From Tables 7.4.18 and 7.3.19 the results reveal that even the misclassification rates may be cancelled out by interaction with data distribution and different orientations between the two predictor variables. However, whether the data consists of outliers or not, and whether the variance-covariance matrices across groups are identical or not, are seen to be key factors in affecting the classification performance in the Logit method.

#### GDR method interaction effects and main effects on testing data

##### 1. Two-way interaction effects

###### (1) Dist by Disp effect

Table 7.4.20 and Figure 7.4.17 present the paired comparisons and interaction effects of Dist by Disp on testing data for the GDR method. As with the training sample, the data with outlying values provide higher misclassification rates than the data without them across all levels of group dispersion. Additionally, the classification performance is again demonstrated to be better in equal group dispersion data than that in an unequal group dispersion situation.

###### (2) Dist by Orien effect

Inspecting the results displayed in Table 7.3.21 and Figure 7.3.18, we find that in terms of orientation schemes, there are significant interaction effects when data is mixed with different distributions. The difference is trivial between skewed and normal distribution. However, significant differences occur in cases with extreme points. Meanwhile, the Orientation I scheme does not always have lower misclassification rates on generalisation. For example, in skewed distribution, the misclassification rate of Orientation I is slightly higher than that of Orientation II.

###### (3) Disp by Orien effect

Table 7.4.22 reports the multiple-comparison results, and Figure 7.4.19 illustrates the effects of Disp by Orien interaction on GDR testing data. Despite the nonlinear discriminating feature in GDR, the evidence for both training and testing results clearly shows that when the group dispersion is heterogeneous and the predictor

variables have low inter-correlation (group dispersion b), GDR performs worse than it does with homogenous data (group dispersion d). This phenomenon is consistent with MDA & Logit, but contradicts the results found in the training samples of Projection methods – the other neural network.

On the other hand, if the predictor variables are highly correlated, the advantage of equality of variance-covariance matrices across groups may be confounded. This implies that the inter-correlation among predictor variables does have implications for classification ability.

### 3. Main effects

The main effects in GDR on testing data have been identified as significant for data distribution and orientation schemes, but not significant for group dispersion levels.

Seeing the multiple comparison outcomes and mean error rates in Table 7.4.23 and Table 7.4.24, we find that only one pairwise comparison is shown to have a significant difference between group dispersion levels (between group dispersion b and group dispersion d). This occurs because the inter-correlation between predictors will not allow for the direct evaluation of the impact of variance-covariance structures. In effect, as is shown, highly correlated predictor variables do not necessarily make the performance of homogeneous variance-covariance data better than that in the cases of *heterogeneous* variance-covariance structure.

## Proj method interaction effects and main effects on testing data

### 1. Two-way interaction effects

#### (1) Dist by Disp effect

Like all the other three techniques, Projection shows a worse performance in the case with outliers for testing data. However, unlike the other three, there is no interaction between skewed distribution and normal distribution as well as skewed distribution and outlier data. Additionally, in terms of four levels of group dispersion, the evidence presented in Table 7.3.25 and Figure 7.3.20 reveals that the inequality of group dispersion is not the cause of a worse classification. On the

contrary, the misclassification in group dispersion c and d is lower than in their competitive case, group dispersion a and b. To further explore the impact of this factor, we will discuss it in the subsequent analysis of Disp by Orien effects.

(2) Dist by Orien effect

The results of multiple comparisons and the interaction effect of Dist by Orien are provided in Table 7.3.26 and graphed in Figure 7.3.21 respectively. Similarly to the training data results, in the Projection method orientations between two indicators always have a significant interaction when the data has a mixture of different distributions. The biggest difference in misclassification rates between Orientation I and Orientation II (both have a linear boundaries tendency) occurs in the data with the presence of outlying values. This phenomenon also happens to the outcomes demonstrated in MDA, Logit and GDR.

(3) Disp by Orien effect

Table 7.3.27 displays the multiple-comparison results and Figure 7.3.22 illustrates the effects of Disp by Orien interaction on Proj testing data. We note that, as was shown in the training data, classification ability is worse for equal group dispersion than for unequal group dispersion. The graph indicates that the misclassifications in group dispersion a (equal group dispersion with high within-group correlation between two attributes) and in group dispersion b (equal group dispersion with low within-group correlation between two attributes) are higher than their competitive cases of group dispersion c and group dispersion d across all orientation schemes. These outcomes not only support the findings in the training data, but also conform to the results just presented on Dist by Disp interaction effects.

### 3. Main effects

The main effects on the Projection algorithm in testing data are reported in Tables 7.4.28 and 7.4.29. For data distribution and orientation scheme, the results have shown that there are significant differences in most pairwise comparisons. This is consistent with the findings in univariate analysis and is also consistent with the results for involved interactions.



Table 7.4.1 Misclassification Summary  
in Normal Distribution Data  
Using Testing Sample Results

Case	MDA	Logit	GDR	Proj
				Type Error
Aa1	22.67	22.83	22.58	19.92 ✓
Aa2	18.83	18.42	18.67	17.00 ✓
Aa3	28.58	27.08 ✓	33.67	30.41
Ab1	16.67	16.25 ✓	16.75	17.25
Ab2	25.08	24.33 ✓	25.58	26.33
Ab3	26.50	25.33 ✓	29.00	30.33
Ac1	34.67	32.67	39.17	32.75 ✓
Ac2	31.58	30.08	32.17	28.83 ✓
Ac3	38.92	37.83	46.17	30.17 ✓
Ad1	31.83	28.58	33.58	28.17 ✓
Ad2	35.17 ✓	32.92	39.42	35.58
Ad3	37.17	36.25	43.67	36.08 ✓
Overall Error				
Aa1	23.33 ✓	25.83	24.50	28.67
Aa2	19.42 ✓	18.83	18.75	21.17
Aa3	21.33 ✓	22.83	15.42	17.33
Ab1	18.50	17.92 ✓	18.17	21.83
Ab2	24.08	24.58	23.50 ✓	25.33
Ab3	24.75	25.42	22.33	22.08 ✓
Ac1	20.42	24.00	14.17	12.75 ✓
Ac2	17.17	20.08	12.58	9.67 ✓
Ac3	21.67	22.83	7.42 ✓	8.67
Ad1	14.67	17.75	11.25	10.92 ✓
Ad2	23.00	24.67	14.75	9.92 ✓
Ad3	23.50	24.75	12.50 ✓	14.17
Overall Error				
Aa1	23.00 ✓	24.33	23.54	24.29
Aa2	19.13	18.62 ✓	18.71	19.08
Aa3	24.96	24.96	24.54	23.87 ✓
Ab1	17.58	17.08 ✓	17.46	19.34
Ab2	24.58	24.16 ✓	24.54	25.83
Ab3	25.63	25.38 ✓	25.67	26.21
Ac1	27.54	28.33	26.96	22.75 ✓
Ac2	24.38	25.08	22.37	19.25 ✓
Ac3	30.29	30.33	26.79	19.42 ✓
Ad1	23.25	23.17	22.42	19.54 ✓
Ad2	29.08	28.79	27.08	22.75 ✓
Ad3	30.33	30.50	28.08	25.13 ✓

Table 7.4.2 Misclassification Summary  
in Skewed Distribution Data  
Using Testing Sample Results

Case	MDA	Logit	GDR	Proj
				Type Error
Ba1	32.58 ✓	33.58	38.83	42.67
Ba2	42.33	41.17	34.25	27.42 ✓
Ba3	12.50	11.58 ✓	12.92	13.67
Bb1	29.58 ✓	30.42	35.83	41.50
Bb2	26.42	26.58	30.08	33.50
Bb3	17.75 ✓	18.58	23.42	25.75
Bc1	40.17	34.92	33.33	35.00 ✓
Bc2	42.08	38.75	38.58	35.17 ✓
Bc3	18.58	14.58 ✓	16.75	16.08
Bd1	37.00	32.25 ✓	32.33	33.25
Bd2	39.25	34.92	34.42	30.25 ✓
Bd3	29.00	26.00 ✓	29.00	28.00
Type II Error				
Ba1	40.17	37.50	30.00	28.25 ✓
Ba2	34.75 ✓	36.17	39.00	45.92
Ba3	9.08	9.75	7.17	5.50 ✓
Bb1	42.42	39.67	32.67	26.92 ✓
Bb2	33.42	32.83	27.25	24.85 ✓
Bb3	26.08	23.75	15.25	11.42 ✓
Bc1	20.17 ✓	24.67	25.92	25.00
Bc2	18.75 ✓	22.67	22.33	22.67
Bc3	5.75	4.08	2.75	2.17 ✓
Bd1	17.75 ✓	21.58	21.67	25.42
Bd2	17.08 ✓	21.75	22.67	20.50
Bd3	10.42	16.00	9.58	8.33 ✓
Overall Error				
Ba1	36.38	35.54	34.42 ✓	35.46
Ba2	38.54	38.67	36.63 ✓	36.67
Ba3	10.79	10.67	10.04	9.58 ✓
Bb1	36.00	35.04	34.25	34.21 ✓
Bb2	29.92	29.71	28.67	29.17 ✓
Bb3	21.92	21.17	19.33	18.58 ✓
Bc1	30.17	29.79	29.63 ✓	30.00
Bc2	30.42	30.71	30.46	28.92 ✓
Bc3	12.17	9.33	9.75	9.13 ✓
Bd1	27.37	26.92 ✓	27.00	29.33
Bd2	28.17	28.33	28.54	25.37 ✓
Bd3	19.17	21.00	19.29	18.17 ✓

Table 7.4.3 Misclassification Summary  
in Symmetric Distribution with Outliers  
Using Testing Sample Results

Case	MDA	Logit	GDR	Proj
				Type Error
Ca1	21.67 ✓	25.83	29.92	33.92
Ca2	34.53	33.42	32.75	32.50 ✓
Ca3	39.25 ✓	37.50	36.92	40.33
Cb1	16.33 ✓	20.17	23.67	29.58
Cb2	39.42	36.75	36.25 ✓	36.67
Cb3	39.08	36.33	35.75	32.08 ✓
Cc1	28.83 ✓	29.58	37.58	43.75
Cc2	43.58	41.08	43.42	42.00 ✓
Cc3	45.08	43.33 ✓	45.25	48.42
Cd1	27.92 ✓	29.08	30.25	39.58
Cd2	40.17	38.42 ✓	39.75	47.42
Cd3	43.67	42.08	43.92	50.25
Overall Error				
Ca1	39.50	31.75	25.42	20.25 ✓
Ca2	31.50 ✓	32.50	33.67	32.08
Ca3	31.58	32.67	32.58	28.50 ✓
Cb1	38.33	32.42	26.00	20.42 ✓
Cb2	28.83	28.41 ✓	30.92	29.08
Cb3	30.17 ✓	31.33	32.25	34.75
Cc1	39.17	37.67	23.25	13.33 ✓
Cc2	26.25	26.00	24.00	19.67 ✓
Cc3	30.75	31.83	25.00	15.67 ✓
Cd1	36.83	34.08	30.25	13.25 ✓
Cd2	29.83	31.50	28.92	18.00 ✓
Cd3	29.75	30.90	28.50	17.58 ✓
Overall Error				
Ca1	30.33	28.79	27.67	27.08 ✓
Ca2	33.02	32.96	33.21	32.29 ✓
Ca3	35.42	35.08	34.75	34.42 ✓
Cb1	27.33	26.29	24.83 ✓	25.00
Cb2	34.13	32.58 ✓	33.58	32.87
Cb3	34.63	33.83	34.00	33.42 ✓
Cc1	34.00	33.63	30.42	28.54 ✓
Cc2	34.92	33.54	33.71	30.83 ✓
Cc3	37.92	37.58	35.13	32.04 ✓
Cd1	30.37	31.58	30.25	26.42 ✓
Cd2	35.00	34.96	34.33	32.71 ✓
Cd3	36.71	36.49	36.21	33.92 ✓

**Table 7.4.4 The Pattern of Type I and Type II Errors for Each of Four Methods**

Case	MDA	Logit	GDR	Proj
Aa1	+	+	+	+
Aa2	+	+	+	+
Aa3	-	-	-	-
Ab1	+	+	+	+
Ab2	-	+	-	-
Ab3	-	+	-	-
Ac1	-	-	-	-
Ac2	-	-	-	-
Ac3	-	-	-	-
Ad1	-	-	-	-
Ad2	-	-	-	-
Ad3	-	-	-	-
Ba1	+	+	+	+
Ba2	-	-	-	-
Ba3	-	-	-	-
Bb1	+	+	+	+
Bb2	+	+	+	+
Bb3	+	+	+	+
Bc1	-	-	-	-
Bc2	-	-	-	-
Bc3	-	-	-	-
Bd1	-	-	-	-
Bd2	-	-	-	-
Bd3	-	-	-	-
Ca1	+	+	+	+
Ca2	-	-	-	-
Ca3	-	-	-	-
Cb1	+	+	+	+
Cb2	-	-	-	-
Cb3	-	-	-	-
Cc1	+	+	+	+
Cc2	-	-	-	-
Cc3	-	-	-	-
Cd1	+	+	+	+
Cd2	-	-	-	-
Cd3	-	-	-	-

**Table 7.4.1 The Change and Percentage Change in Overall Errors between GDR, Proj and STMs**

Case	GDR		STMs Method	
	Change in	%Change	Proj	STMs Method
Aa1	-0.13	-0.53	0.63	2.64
Aa2	-0.17	-0.87	0.21	1.09
Aa3	-0.42	-1.68	-1.09	-4.37
Ab1	0.13	0.75	2.21	12.75
Ab2	0.17	0.70	1.46	5.99
Ab3	0.17	0.65	0.71	2.76
Ac1	-0.98	-3.49	-5.19	-18.56
Ac2	-2.36	-9.54	-5.48	-22.16
Ac3	-3.52	-11.63	-10.89	-35.93
Ad1	-0.79	-3.40	-3.67	-15.81
Ad2	-1.86	-6.41	-6.19	-21.38
Ad3	-2.34	-7.68	-5.29	-17.38
Ba1	-1.54	-4.28	-0.5	-1.39
Ba2	-1.98	-5.12	-1.94	-5.01
Ba3	-0.69	-2.43	-1.15	-10.72
Bb1	-1.27	-3.58	-1.31	-3.69
Bb2	-1.15	-3.84	-0.65	-2.16
Bb3	-2.22	-10.28	-2.97	-13.76
Bc1	-0.35	-1.17	0.03	0.07
Bc2	-0.11	-0.34	-1.65	-5.38
Bc3	-1.00	-9.30	-1.62	-15.07
Bd1	-0.15	-0.53	2.19	8.05
Bd2	0.29	1.03	-2.88	-10.20
Bd3	-0.80	-3.96	-1.92	-9.53
Ca1	-1.89	-6.39	-2.48	-8.39
Ca2	0.22	0.67	-0.70	-2.12
Ca3	0.50	-1.42	-0.83	-2.36
Cb1	-1.98	-7.39	-1.81	-6.75
Cb2	0.23	0.68	-0.49	-1.45
Cb3	-0.23	-0.67	-0.81	-2.37
Cc1	-3.40	-10.0	-3.28	-15.60
Cc2	-0.52	-1.52	-3.40	-9.93
Cc3	-2.62	-6.94	-5.71	-15.13
Cd1	-1.73	-5.40	-5.56	-17.37
Cd2	-0.65	-1.86	-2.27	-6.49
Cd3	-0.39	-1.07	-2.68	-7.32

+ denotes the situation in which Type II error is larger than Type I error  
- denotes the situation in which Type II error is lower than Type I error  
• denotes the situation in which Type II error equals Type I error

**Table 7.4.2 The Change and Percentage Change in Type I, Type II Errors between GDR and STMs**

Case	Change in Type I	%Change in Type I	Change in Type II	%Change in Type II	-in Type I	-in Type II	Both
Aa1	-0.17	-0.08	-0.75	0.33	v		
Aa2	0.05	-0.38	0.24	-1.96		v	
Aa3	5.84	-6.66	20.99	-30.16		v	
Ab1	0.29	-0.04	1.76	-0.22		v	
Ab2	0.88	-0.83	3.54	-3.41		v	
Ab3	3.09	-2.76	11.90	-10.98		v	
Ac1	5.50	-8.04	16.34	-36.20		v	
Ac2	1.34	-6.05	4.35	-32.46		v	
Ac3	7.80	-14.83	20.31	-66.65		v	
Ad1	3.38	-4.96	11.17	-30.60		v	
Ad2	5.38	-9.09	15.79	-38.12		v	
Ad3	6.96	-11.63	18.96	-48.19		v	
Ba1	5.75	-8.84	17.38	-22.75		v	
Ba2	-7.50	3.54	-17.96	9.98	v		
Ba3	0.88	-2.25	7.31	-23.85		v	
Bb1	5.83	-8.38	19.43	-20.40		v	
Bb2	3.58	-5.88	13.51	-17.74		v	
Bb3	5.26	-9.66	28.93	-38.79		v	
Bc1	-4.22	3.50	-11.23	15.61	v		
Bc2	-1.84	1.62	-4.54	7.82		v	
Bc3	0.17	-2.17	1.03	-44.05		v	
Bd1	-2.30	2.01	-6.63	10.20		v	
Bd2	-2.67	3.26	-7.19	16.77		v	
Bd3	1.50	-3.63	5.46	-27.48		v	
Ca1	6.17	-10.21	25.98	-28.65		v	
Ca2	-1.23	1.67	-3.61	5.22		v	
Ca3	-1.46	0.46	-3.79	1.42		v	
Cb1	5.42	-9.38	29.70	-26.50		v	
Cb2	-1.84	2.30	-4.82	8.04		v	
Cb3	-1.96	1.50	-5.19	4.88		v	
Cc1	8.38	-15.17	28.68	-39.49		v	
Cc2	1.09	-2.13	2.58	-8.13		v	
Cc3	1.05	-6.29	2.36	-20.10		v	
Cd1	1.75	-5.21	6.14	-14.68		v	
Cd2	0.46	-1.75	1.16	-5.69		v	
Cd3	1.05	-1.83	2.44	-6.02		v	

**Table 7.4.3 The Change and Percentage Change in Type I, Type II Errors between GDR and STMs**

Case	Change in Type I	%Change in Type I	Change in Type II	%Change in Type II	-in Type I	-in Type II	Both
Aa1	-2.83	4.09	-12.44	16.64		v	
Aa2	-1.63	2.05	-8.73	10.69		v	
Aa3	2.58	-4.75	9.27	-21.51		v	
Ab1	0.79	3.62	4.80	19.88			
Ab2	1.63	1.00	6.58	4.11			
Ab3	4.42	-3.01	17.04	-11.98		v	
Ac1	-0.92	-9.46	-2.73	-42.59		v	
Ac2	-2.00	-8.96	-6.49	-48.08		v	
Ac3	-8.21	-13.58	-21.38	-61.03		v	
Ad1	-2.04	-5.29	-6.74	-32.63		v	
Ad2	1.54	-13.92	4.51	-58.38		v	
Ad3	-6.30	-9.96	-1.72	-41.26		v	
Ba1	9.59	-10.59	28.99	-27.26		v	
Ba2	-14.33	10.46	-34.32	29.50		v	
Ba3	1.63	-3.92	13.54	-41.58		v	
Bb1	11.50	-14.13	38.33	-34.41		v	
Bb2	7.00	-8.30	26.42	-25.04		v	
Bb3	7.59	-13.50	41.76	-54.16		v	
Bc1	-2.55	2.58	-6.78	11.51		v	
Bc2	-5.25	1.96	-12.98	9.46		v	
Bc3	-0.50	-2.75	-3.02	-55.85		v	
Bd1	-1.38	5.76	-3.97	29.27		v	
Bd2	-6.84	1.09	-18.43	5.59		v	
Bd3	0.50	-4.88	1.82	-36.94		v	
Ca1	10.17	-15.38	42.82	-43.16		v	
Ca2	-1.48	0.08	-4.34	0.25		v	
Ca3	1.96	-3.63	5.09	-11.28		v	
Cb1	11.33	-14.96	62.08	-42.28		v	
Cb2	-1.42	0.46	-3.72	1.61		v	
Cb3	-5.63	4.00	-14.92	13.01		v	
Cc1	14.55	-25.09	49.80	-65.31		v	
Cc2	-0.33	-6.46	-0.78	-24.71		v	
Cc3	4.22	-15.62	9.54	-49.92		v	
Cd1	11.08	-22.21	38.88	-62.63		v	
Cd2	8.13	-12.67	20.68	-41.30		v	
Cd3	7.38	-12.75	17.20	-42.03		v	

Table 7.4.5 Multivariate Analysis of Variance of MDA, Logit, GDR and Proj Methods on Overall Error Rates Using Testing Data

	Wilks'	Hypoth. df	Error of	Multivariate
Dist	0.48	8	1,362	76.601*
Disp	0.83	12	1802.48	10.9303*
Orien	0.8	8	1362	20.3144*
Dist by Disp	0.77	16	2376.935	7.6872*
Dist by Orien	0.35	24	2081.127	53.6766*
Disp by Orien	0.84	24	2376.935	5.0457*
Dist by Disp by Orien	0.71	48	2625.322	5.0979*

Table 7.4.6 Univariate Analysis of Variance for MDA Using Testing Samples

Source of Variance	Sum of Square	DF	Mean Square of	F Value
Dist	0.48	8	1,362	76.601*
Disp	0.83	12	1802.48	10.9303
Orien	0.8	8	1362	20.3144
Dist by Disp	0.77	16	2376.93	7.6872*
Dist by Orien	0.35	24	2081.12	53.6766
Disp by Orien	0.84	24	2376.93	5.0457*
Dist by Disp by	0.71	48	2625.32	5.0979*

Table 7.4.7 Univariate Analysis of Variance for Logit Using Testing Samples

Source of Variance	Sum of Square	DF	Mean Square of	F Value
Dist	8885.194	2	4442.597	206.32*
Disp	376.18	3	125.39	5.82*
Orien	1487.210	2	743.61	34.53*
Dist by Disp	3052.963	6	508.83	23.63*
Dist by Orien	14903.20	4	3725.801	173.03*
Disp by Orien	1508.506	6	251.42	11.68*
Dist by Disp by	3636.383	12	303.03	14.07*

Table 7.4.8 Univariate Analysis of Variance for GDR Using Testing Samples

Source of Variance	Sum of Square	DF	Mean Square of	F Value
Dist	9327.455	2	4663.727	255.97*
Disp	160.58	3	53.53	
Orien	1941.418	2	970.71	53.28*
Dist by Disp	1417.270	6	236.21	12.96*
Dist by Orien	15407.00	4	3851.750	211.4*
Disp by Orien	1243.346	6	207.22	11.37*
Dist by Disp by	2912.712	12	242.73	13.32*

Table 7.4.9 Univariate Analysis of Variance for Proj Using Testing Samples

Source of Variance	Sum of Square	DF	Mean Square of	F Value
Dist	8867.532	2	4433.766	272.19*
Disp	796.1	3	265.37	16.29*
Orien	2412.262	2	1206.131	74.05*
Dist by Disp	524.17	6	87.36	5.36*
Dist by Orien	15940.62	4	3985.157	244.65*
Disp by Orien	1510.930	6	251.82	15.46*
Dist by Disp by	2447.065	12	203.92	12.52*

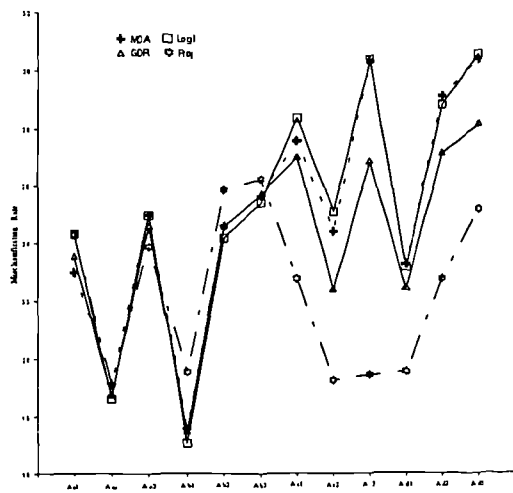


Figure 7.4.1 Comparison on Normal Distribution for Testing Data

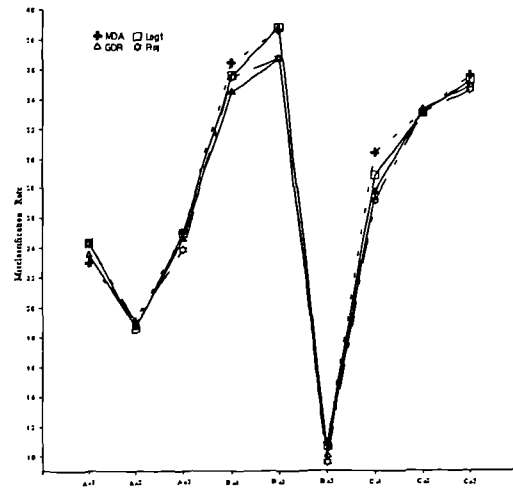


Figure 7.4.4 Comparison on Group Dispersion (a) for Testing Data

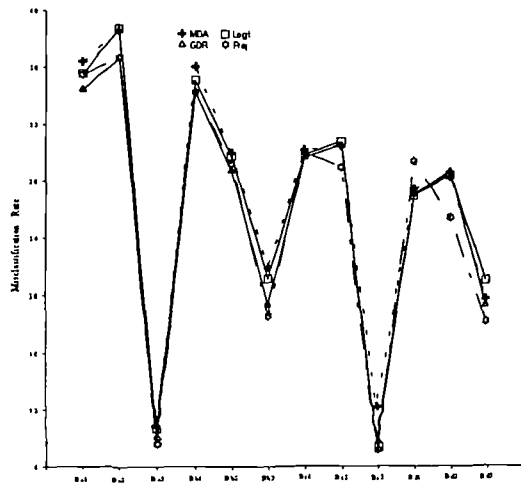


Figure 7.4.2 Comparison on Skewed Distribution for Testing Data

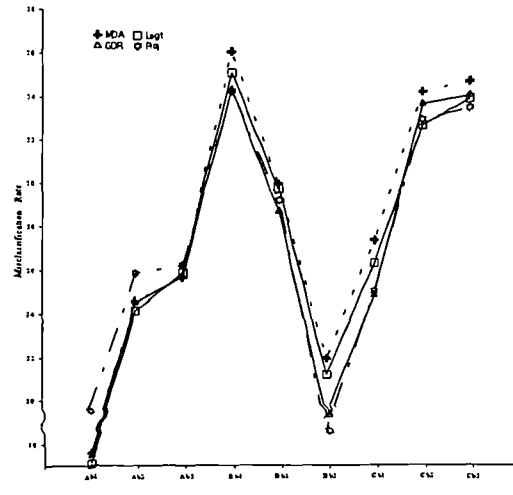


Figure 7.4.5 Comparison on Group Dispersion (b) for Testing Data

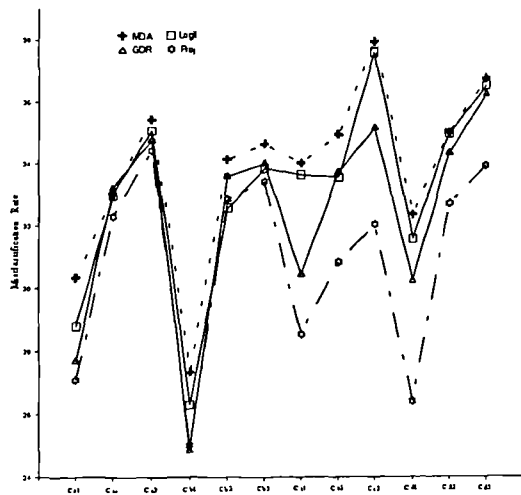


Figure 7.4.3 Comparison on Symmetric Distribution with Outliers for Testing Data

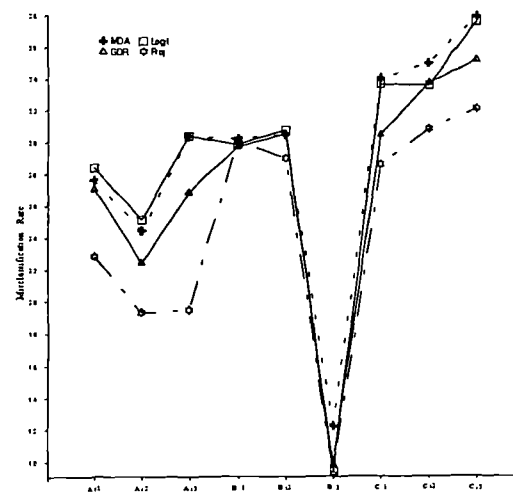


Figure 7.4.6 Comparison on Group Dispersion (c) for Testing Data

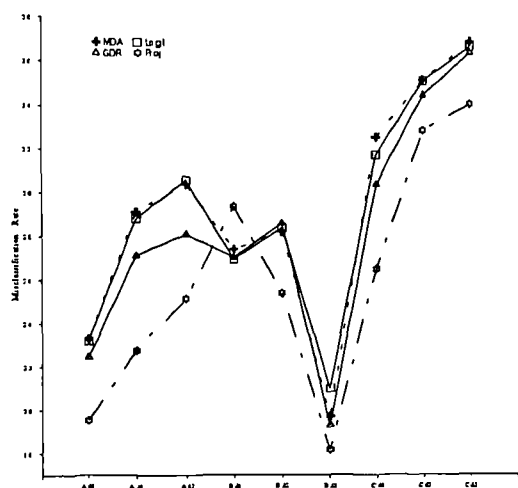


Figure 7.4.7 Comparison on Group Dispersion (d) for Testing Data

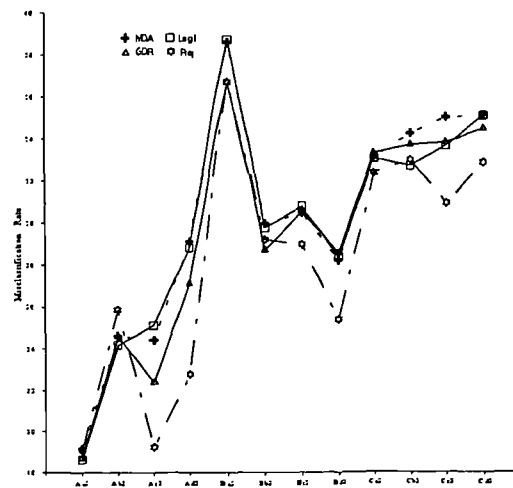


Figure 7.4.9 Comparison on Orientation Scheme II for Testing Data

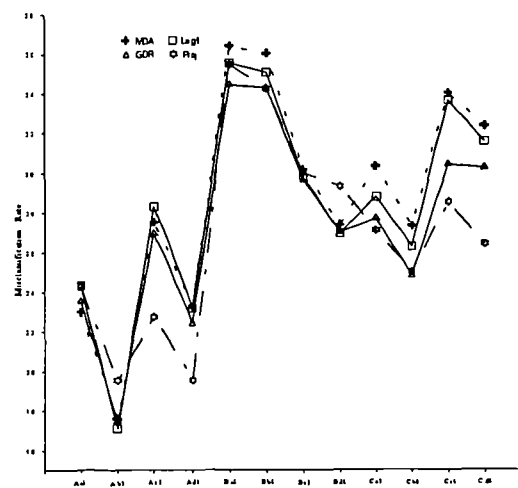


Figure 7.4.8 Comparison on Orientation Scheme I for Testing Data

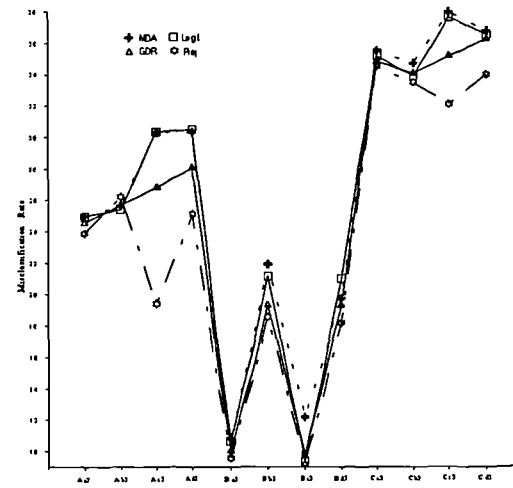


Figure 7.4.10 Comparison on Orientation Scheme III for Testing Data

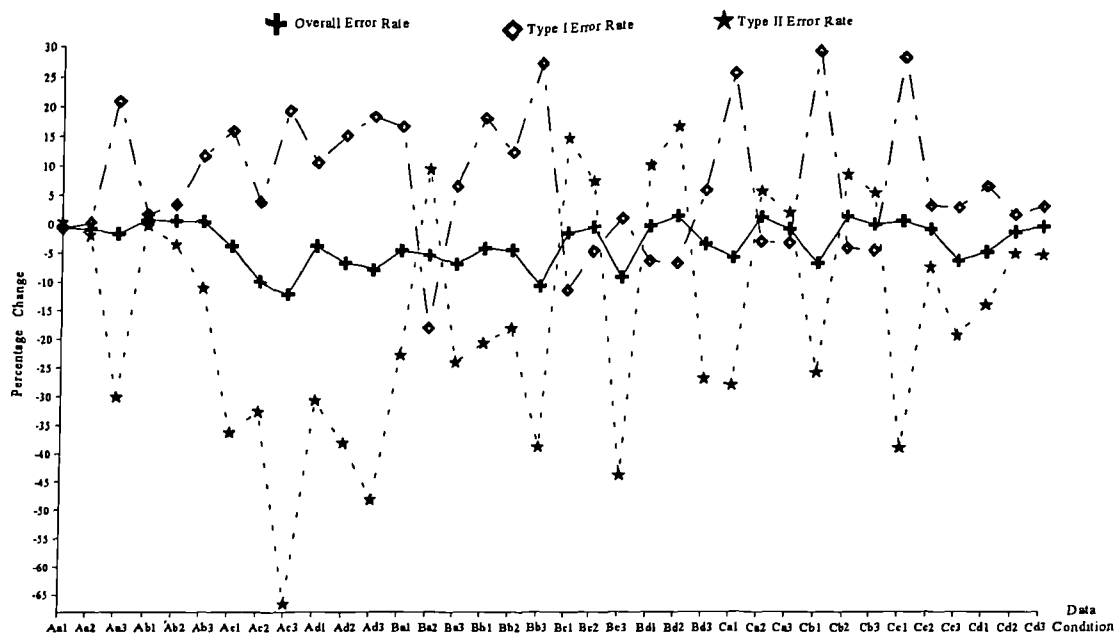


Figure 7.4.I The Plot of Percentage Change in Overall, Type I and Type II Error Rates

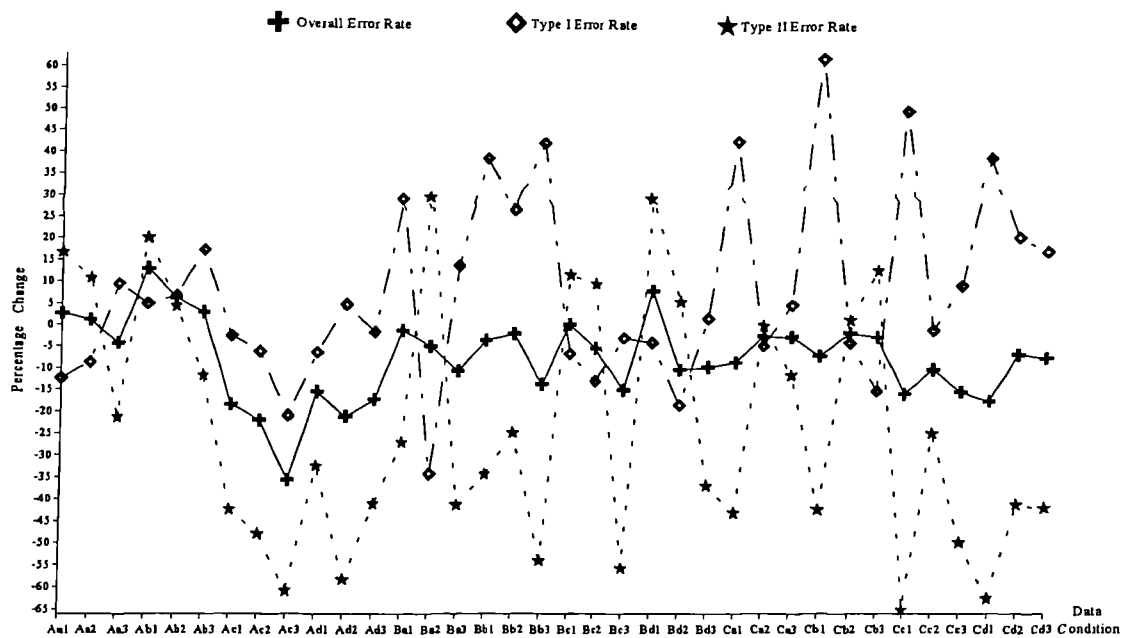


Figure 7.4.II The Plot of Percentage Change in Overall, Type I and Type II Error Rates  
between Proj and Statistical Methods Using Testing Samples

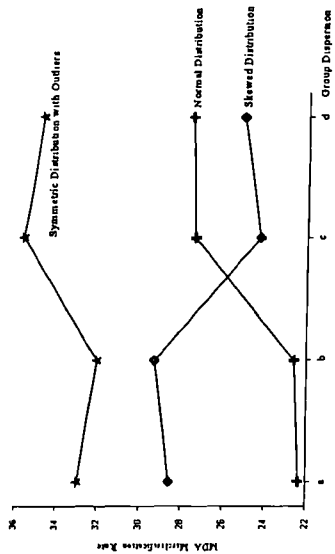


Figure 7.4.11 Interaction Effects of Dist by Disp on MDA  
Using Testing Samples

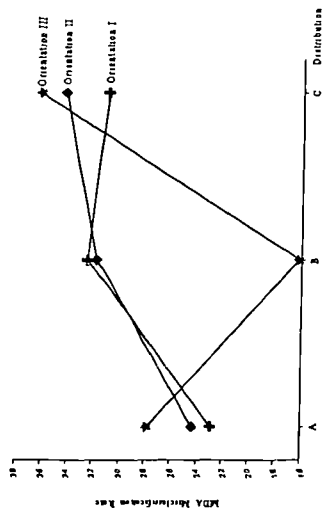


Figure 7.4.12 Interaction Effects of Dist by Orien on MDA  
Using Testing Samples

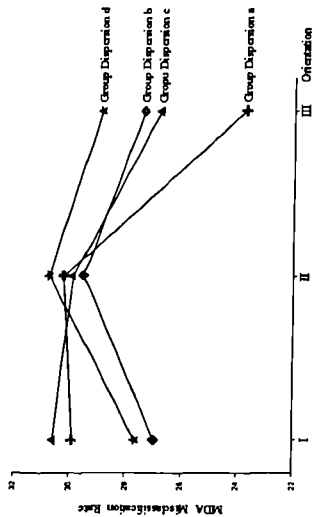


Figure 7.4.13 Interaction Effects of Disp by Orien on MDA  
Using Testing Samples

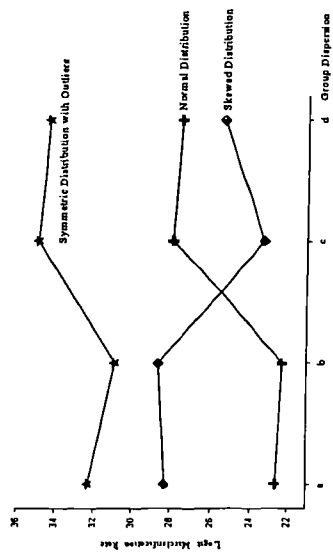


Figure 7.4.14 Interaction Effects of Dist by Orien on Logit  
Using Testing Samples

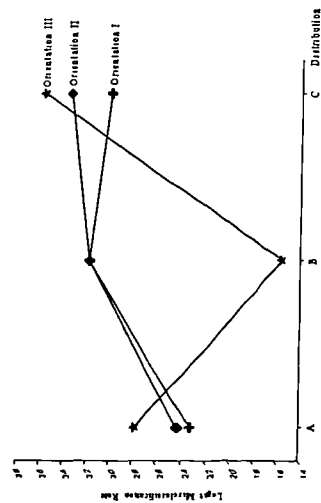


Figure 7.4.15 Interaction Effects of Dist by Orien on Logit  
Using Testing Samples

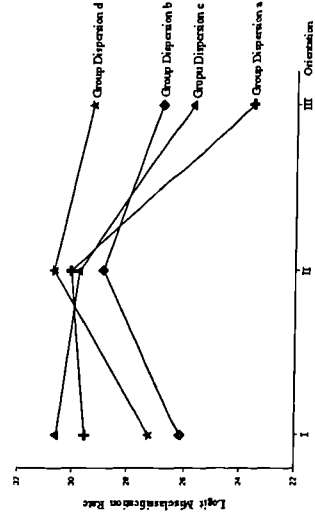


Figure 7.4.16 Interaction Effects of Dist by Orien on Logit  
Using Testing Samples

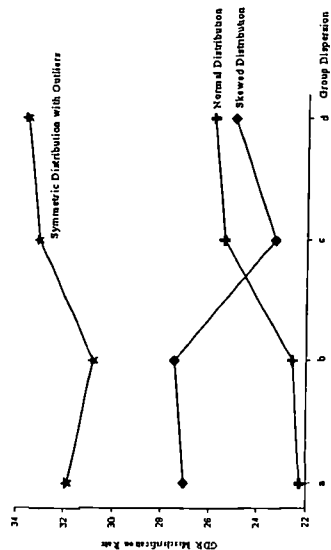


Figure 7.4.17 Interaction Effects of Dist by Disp on GDR  
Using Testing Samples

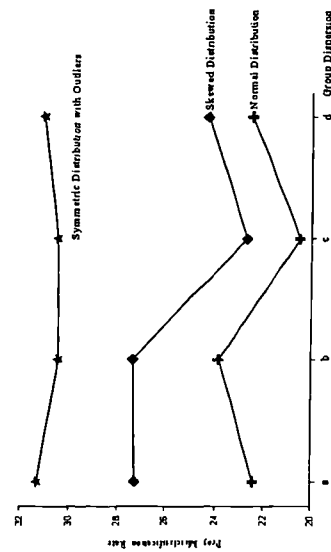


Figure 7.4.20 Interaction Effects of Dist by Disp on Proj  
Using Testing Samples

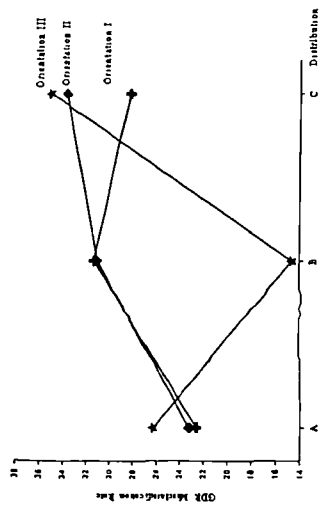


Figure 7.4.18 Interaction Effects of Dist by Orien on GDR  
Using Training Samples

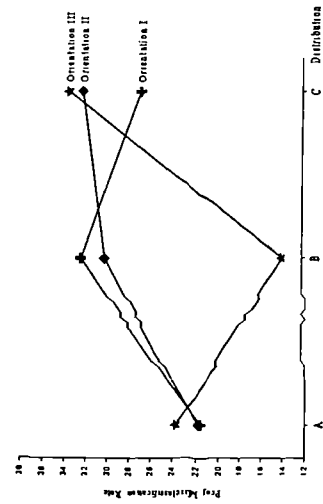


Figure 7.4.21 Interaction Effects of Dist by Orien on Proj  
Using Training Samples

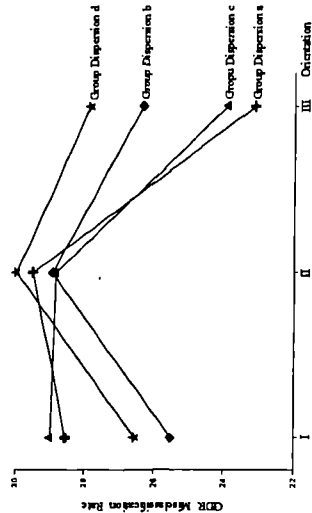


Figure 7.4.19 Interaction Effects of Dist by Orien on GDR  
Using Testing Samples

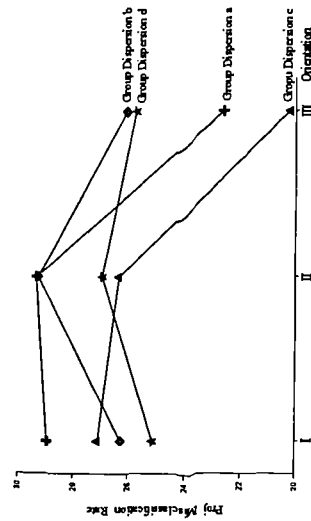


Figure 7.4.22 Interaction Effects of Dist by Orien on Proj  
Using Testing Samples



Table 7.4.10 Multiple Comparison on MDA  
Dist by Disp Effect for Testing Data

Aa	Ab	Ac	Ad	Ba	Bb	Bc	Bd	Ca	Cb	Cc	Cd
Aa	—	*	*	*	*	*	*	*	*	*	*
Ab	—	*	*	*	*	*	*	*	*	*	*
Ac	—	*	*	*	*	*	*	*	*	*	*
Ad	—	*	*	*	*	*	*	*	*	*	*
Ba	—	*	*	*	*	*	*	*	*	*	*
Bb	—	*	*	*	*	*	*	*	*	*	*
Bc	—	*	*	*	*	*	*	*	*	*	*
Bd	—	*	*	*	*	*	*	*	*	*	*
Ca	—	*	*	*	*	*	*	*	*	*	*
Cb	—	*	*	*	*	*	*	*	*	*	*
Cc	—	*	*	*	*	*	*	*	*	*	*
Cd	—	*	*	*	*	*	*	*	*	*	*

Table 7.4.11 Multiple Comparison  
on MDA Dist by Orien Effect  
for Testing Data

A1	A2	A3	B1	B2	B3	C1	C2	C3
A1	—	*	*	*	*	*	*	*
A2	—	*	*	*	*	*	*	*
A3	—	*	*	*	*	*	*	*
B1	—	*	*	*	*	*	*	*
B2	—	*	*	*	*	*	*	*
B3	—	*	*	*	*	*	*	*
C1	—	*	*	*	*	*	*	*
C2	—	*	*	*	*	*	*	*
C3	—	*	*	*	*	*	*	*

Table 7.4.12 Multiple Comparison on MDA  
Dist by Orien Effect for Testing Data

a1	a2	a3	b1	b2	b3	c1	c2	c3	d1	d2	d3
a1	—	*	*	*	*	*	*	*	*	*	*
a2	—	*	*	*	*	*	*	*	*	*	*
a3	—	*	*	*	*	*	*	*	*	*	*
b1	—	*	*	*	*	*	*	*	*	*	*
b2	—	*	*	*	*	*	*	*	*	*	*
b3	—	*	*	*	*	*	*	*	*	*	*
c1	—	*	*	*	*	*	*	*	*	*	*
c2	—	*	*	*	*	*	*	*	*	*	*
c3	—	*	*	*	*	*	*	*	*	*	*
d1	—	*	*	*	*	*	*	*	*	*	*
d2	—	*	*	*	*	*	*	*	*	*	*
d3	—	*	*	*	*	*	*	*	*	*	*

Table 7.4.14 Mean Error on MDA for Testing Data

Le	Mean Error Rate	Lc	Mean Error Rate	Leve	Mean Error Rate
A	24.98		27.95		28.78
B	26.8		27.97		30.11
C	33.81		29.09		26.71
			29.110		

Table 7.4.15 Multiple Comparison on Logit  
Dist by Disp Effect for Testing Data

Aa	Ab	Ac	Ad	Ba	Bb	Bc	Bd	Ca	Cb	Cc	Cd
Aa	—	*	*	*	*	*	*	*	*	*	*
Ab	—	*	*	*	*	*	*	*	*	*	*
Ac	—	*	*	*	*	*	*	*	*	*	*
Ad	—	*	*	*	*	*	*	*	*	*	*
Ba	—	*	*	*	*	*	*	*	*	*	*
Bb	—	*	*	*	*	*	*	*	*	*	*
Bc	—	*	*	*	*	*	*	*	*	*	*
Bd	—	*	*	*	*	*	*	*	*	*	*
Ca	—	*	*	*	*	*	*	*	*	*	*
Cb	—	*	*	*	*	*	*	*	*	*	*
Cc	—	*	*	*	*	*	*	*	*	*	*
Cd	—	*	*	*	*	*	*	*	*	*	*

Table 7.4.17 Multiple Comparison on Logit  
Dist by Orien Effect for Testing Data

a1	a2	a3	b1	b2	b3	c1	c2	c3	d1	d2	d3
a1	—	*	*	*	*	*	*	*	*	*	*
a2	—	*	*	*	*	*	*	*	*	*	*
a3	—	*	*	*	*	*	*	*	*	*	*
b1	—	*	*	*	*	*	*	*	*	*	*
b2	—	*	*	*	*	*	*	*	*	*	*
b3	—	*	*	*	*	*	*	*	*	*	*
c1	—	*	*	*	*	*	*	*	*	*	*
c2	—	*	*	*	*	*	*	*	*	*	*
c3	—	*	*	*	*	*	*	*	*	*	*
d1	—	*	*	*	*	*	*	*	*	*	*
d2	—	*	*	*	*	*	*	*	*	*	*
d3	—	*	*	*	*	*	*	*	*	*	*

Table 7.4.18 Main Effects on Logit for Testing Data

A	B	C	a	b	c	d	I	II	III
A	—	*	*	*	*	*	I	—	*
B	—	*	*	*	*	*	II	—	*
C	—	*	*	*	*	*	III	—	*

Table 7.3.19 Main Error on Logit for Testing Data

Le	Mean Error Rate	Lc	Mean Error Rate	Leve	Mean Error Rate
A	25.09	a	27.74	I	28.38
B	26.41	b	27.28	II	29.87
C	33.11	c	28.7	III	26.36
		d	29.08		

\* denotes significant using a 0.05 family-wise alpha

Table 7.4.20 Multiple Comparison on GDR  
Dist by Disp Effect for Testing Data

Aa	Ab	Ac	Ad	Ba	Bb	Bc	Bd	Ca	Cb	Cc	Cd
Aa	—			*	*	*	*	*	*	*	*
Ab	—			*	*	*	*	*	*	*	*
Ac		—		*	*	*	*	*	*	*	*
Ad			—	*	*	*	*	*	*	*	*
Ba				—	*	*	*	*	*	*	*
Bb					—	*	*	*	*	*	*
Bc						—	*	*	*	*	*
Bd							—	*	*	*	*
Ca								—	*	*	*
Cb									—	*	*
Cc										—	*
Cd											—

Table 7.4.21 Multiple Comparison  
on GDR Dist by Orien Effect  
for Testing Data

A1	A2	A3	B1	B2	B3	C1	C2	C3
A1	—	*	*	*	*	*	*	*
A2	—	*	*	*	*	*	*	*
A3		—	*	*	*	*	*	*
B1			—	*	*	*	*	*
B2				—	*	*	*	*
B3					—	*	*	*
C1						—	*	*
C2							—	*
C3								—

Table 7.4.22 Multiple Comparison on GDR  
Dist by Orien Effect for Testing Data

a1	a2	a3	b1	b2	b3	c1	c2	c3	d1	d2	d3
a1	—	*						*			
a2	—										
a3		—	*	*	*	*	*	*	*	*	*
b1			—	*	*	*	*	*	*	*	*
b2				—	*	*	*	*	*	*	*
b3					—	*	*	*	*	*	*
c1						—	*	*	*	*	*
c2							—	*	*	*	*
c3								—	*	*	*
d1									—	*	*
d2										—	*
d3											—

Table 7.4.23 Main Effects on GDR or Testing Data

A	B	C	a	b	c	d	I	II	III
A	—	*	*	*	*	*	I	—	*
B	—	*	*	*	*	*	II	—	*
C	—	*	*	*	*	*	III	—	*

Table 7.4.24 Main Error on GDR for Testing Data

Le	Mean Error Rate	Le	Mean Error Rate	Leve	Mean Error Rate
A	24.01	a	27.06	I	27.4
B	25.67	b	26.93	II	29.32
C	32.340	c	27.25	III	29.3
		d	28.13		

Table 7.4.25 Multiple Comparison on Proj  
Dist by Disp Effect for Testing Data

Aa	Ab	Ac	Ad	Ba	Bb	Bc	Bd	Ca	Cb	Cc	Cd
Aa	—			*	*	*	*	*	*	*	*
Ab	—			*	*	*	*	*	*	*	*
Ac		—		*	*	*	*	*	*	*	*
Ad			—	*	*	*	*	*	*	*	*
Ba				—	*	*	*	*	*	*	*
Bb					—	*	*	*	*	*	*
Bc						—	*	*	*	*	*
Bd							—	*	*	*	*
Ca								—	*	*	*
Cb									—	*	*
Cc										—	*
Cd											—

Table 7.4.27 Multiple Comparison on Proj  
Dist by Orien Effect for Testing Data

a1	a2	a3	b1	b2	b3	c1	c2	c3	d1	d2	d3
a1	—	*						*			
a2	—	*									
a3		—	*	*	*	*	*	*	*	*	*
b1			—	*	*	*	*	*	*	*	*
b2				—	*	*	*	*	*	*	*
b3					—	*	*	*	*	*	*
c1						—	*	*	*	*	*
c2							—	*	*	*	*
c3								—	*	*	*
d1									—	*	*
d2										—	*
d3											—

Table 7.4.28 Main Effects on Proj for Testing Data

A	B	C	a	b	c	d	I	II	III
A	—	*	*	*	*	*	I	—	*
B	—	*	*	*	*	*	II	—	*
C	—	*	*	*	*	*	III	—	*

Table 7.4.29 Mean Error on Proj for Testing Data

Le	Mean Error Rate	Le	Mean Error Rate	Leve	Mean Error Rate
A	22.31	a	26.97	I	26.85
B	25.38	b	27.2	II	27.98
C	30.8	c	24.54	III	23.66
		d	25.93		

\* denotes significant using a 0.05 family-wise alpha

## 7.5 Summary and Conclusions

This chapter has analysed the results on both training and testing samples under a wide variety of data conditions for four alternative methods. Some important conclusions are summarised as follows

1. Considering the type of misclassification, MDA and Logit appeared to have the same patterns of Type I and Type II error rate. By contrast GDR and Projection neural networks always provided identical direction of the two type errors. However there were many opposite signs between these two classes of techniques, especially in the skewed distribution and the data with extreme points, which are the usual cases in practice. The logic of discriminating methodology between them is shown to be different under some circumstances.
2. In terms of overall classification accuracy, the Projection neural network provided the best performance in learning ability of all four methods over all ranges of data situations. At the same time, the ANN approaches were proved to be overwhelmingly superior to statistical methods in the training phase. However, according to the testing sample results, MDA and Logit have achieved a more successful generalisation ability relative to GDR and Proj. Hence, although ANNs still outperformed traditional statistical methods in most cases, overfitting problem has emerged as a potential obstacle to practitioners.
3. Despite the fact that ANNs produced a better overall performance, it did not necessarily lead to improvement in both Type I and Type II errors. Conversely, the Type I error often increased with a decrease in the Overall error rate. This outcome is particularly obvious in GDR under a normal model. If greater Type I error cost is expected, the conclusion that ANNs outperform statistical methods can not be agreed unanimously in bankruptcy prediction. The use of ANNs or statistical methods probably depends upon the user's purpose or the data situation.
4. With respect to the discriminating capacity of individual methods, there is no strong evidence that the skewed-distribution data caused an impairment of the classification accuracy. However, we have no doubt that outliers affected all techniques adversely. Nonetheless, the ANNs were more robust to these situations. This finding suggests the ANN solutions are more useful for data in which aberrant values or outliers are present.

5. For MDA, Logit and GDR, the results indicated that equal variance-covariance matrices across groups are an important factor in improving classification performance, but this superiority is more evident when the predictor variables have a low inter-correlation. If the predictor variables are highly correlated, the advantage of homogeneity of variance-covariance structures could be confounded due to the interaction effects. Put another way, the correlation between predictor variables, as we pointed out in Chapter Five, sometimes plays a vital role in affecting predictive power.
6. Interestingly and surprisingly, it was found that in terms of the Projection method, classification ability is worse for equal group dispersion than for unequal group dispersion cases. This outcome was presented both in the training data and testing data. However, the reason for this phenomenon is unclear. Whether this extraordinary result can be applied to  $n$  dimension or not is noteworthy and needs further examination.
7. ANNs were proved to be more capable of capturing complex decision frontiers. The ability of multi-layered Perceptron (MLP) to cope with nonlinear boundaries, demonstrated earlier in theoretical terms, is confirmed practically in our study. This is helpful for resolving bankruptcy prediction problem whose decision set is believed to be modelled as a nonlinear form by financial ratios.

The results of this simulation study have provided us with a insight into the weaknesses and strengths of each of alternative methods. Although we have tried to manipulate reasonably comprehensive data conditions in order to test the individual techniques, the results should not be unconditionally interpreted to the more complex data. To verify this result and to apply these techniques in realistic practice, a real financial data analysis will be conducted. Before proceeding with the empirical study, we will discuss the relevant problems encountered in bankruptcy prediction models in the Chapter Eight.

## **Chapter Eight**

# **THE PROBLEMS OF BANKRUPTCY PREDICTION AND THEIR PROPOSED SOLUTIONS**

### **8.1 Introduction**

The results of the simulation study in Chapter Seven have provided some insights into the strengths and weaknesses of the alternative discriminating techniques, and have shown the conditions under which one method is superior to the others. In order to verify the simulation results of the four models in actual practice, a study on real financial data must be conducted. At the same time, in predicting business failure, there are some alternative ways of constructing a model. Therefore, before we perform this empirical study, these problems need to be discussed and the solutions need to be determined in order to guarantee the relevant experiments proceed smoothly and correctly.

The following four problems may affect the usefulness of any empirical study's classification accuracy in making assessments on bankruptcy prediction. They are

1. The selection of predictor variables (financial ratios).
2. Choice-based sampling design.
3. Unequal misclassification costs of Type I and Type II errors
4. Model validation and generalisation.

This chapter will discuss each of these problems and propose solutions to cope with them. These solutions will then be used in the research designs of empirical experiments presented in the next chapter.

### **8.2 Problems with the Selection of Predictor Variables**

One common criticism of bankruptcy prediction models is that they lack theory to guide the selection of predictor variables [Foster, 1986]; [Jones, 1987]. Most studies using financial

ratios have been based on popularity in the literature, potential relevance and subjective judgement [Barnes, 1987]. Because of the absence of a generally accepted theory of financial distress, it is difficult to justify the use of any particular predictor variables. It can be found that a wide variety of financial ratios have been employed in empirical failure predictions. Kares and Prakash [1987] presented a summary table of financial ratios used in previous studies as predictors of failure. It is not surprising that a diverse selection of financial ratios has been used given the limited theoretical basis for choosing the ratios.

This section first discusses the application of previous accrual-based and cash-based financial ratio. Some bankruptcy theories are then presented in order to help us increase the understanding of the behaviour and process of business failure. Finally, a rationale for the selection of predictor variables in this empirical study will be proposed.

### **8.2.1 Accrual Accounting Ratios and Cash Flow Ratios**

Accrual accounting predictors were widely adopted in the early business distress prediction. On the other hand, cash flow or cash position ratios were reported relevant to predict corporate bankruptcy.

Early research selecting accrual-based ratios tended to be ad hoc in order to satisfy the acceptable degree of classification accuracy [Beaver, 1966, 1968], [Altman, 1968], [Deakin, 1972], [Pinches, et al. 1973, 1975] etc. However, they brought into the concept of time series and industrial impact as vertical or horizontal perspective of indicators [Edmister, 1972], [Blum, 1974]. Market risk was also introduced as a consideration for the selection of broadening variables [Beaver, et al. 1970]. This joint effort more or less provides us some directions for this disputable problem.

Studies into the usefulness of cash-based measures as predictors have mostly shown that such measures may not be as effective as expected in distinguishing failing and nonfailing firms [Casey and Bartczak, 1984, 1985], [Gentry et al. 1985a, 1985b], [Gombola et al., 1987], [Aziz and Lawson, 1989]. Researchers were reluctant to agree that failed and nonfailed companies exhibit some clearly differentiated statistical values for various cash-based measures. Nevertheless the decomposition of all cash flows into component parts or their new usage may deserve further attention.

No matter what kind of accounting variables were employed, financial ratios have rarely been used to test hypotheses and theories of economic and financial behaviour [Laitinen, 1991]. Zavgren [1983] stated that in the absence of a theory that indicates the important dimensions, the selection of various closely related ratios in the model may lead to sample-specific results and to the instability of the prediction model.

### **8.2.2 Theoretical Models of Bankruptcy**

Wilcox's study [1971a] is one of the earliest and most primitive theoretical models of bankruptcy. The gambler's ruin model proposed by him assumed that a firm's behaviour is equivalent to a gambler who has a given amount of capital  $K$  that will either grow or be reduced to zero by a series of independent trials. The given capital,  $K$ , of a firm will change randomly and become bankrupt when its worth falls to zero. The change in  $K$  results from cash flows from the firm's operation. Positive change increases  $K$  and negative change requires the firm to liquidate its assets. When a company's  $K$  becomes negative, it is declared bankrupt. The expected probability of bankruptcy as well as the time to reach bankruptcy all just like the probability of the risk in a gamblers' game. Although the theory provided a functional form for the probability of ultimate ruin, the outcome in practical bankruptcy prediction of applying this concept has been disappointing, as it assumes that periodic cash flows are independent of each other. In fact Wilcox [1976] found that most of his sample's data violated the theory's assumptions and discarded the functional structure of this theory. He suggested building a model with variables that demonstrated proven successful classification accuracy in the literature. Furthermore, Santomero and Vinso [1977] produced a grossly inaccurate and implausibly low probability of failure in banking data using Wilcox's model.

Scott [1976, 1977, 1981] pointed out that the Wilcox model is too simplistic and inappropriate to explain the bankruptcy phenomenon, and attempted to improve on this simple model. In his early study, Scott assumed that a company has a potentially infinite life and can meet losses by selling debt or equity in an efficient market without incurring a flotation cost. A firm would remain solvent as long as stockholder wealth measured by market value is positive. Scott later revised the earlier model by introducing assumptions

about the firm's flotation costs because of imperfect access to the external the capital market. He also assumed that there may be a tax system which favours internally-financed corporate investments. According to this modified model, a firm will go bankrupt when the market value of its securities is less than the amount of investment needed at times of negative income. This model implies that bankruptcy is not due to insufficiency of profit but rather to investment errors. However, regardless of the truth or falsity of the implication, the model can still not provide clear guidelines for selecting financial ratios.

A more specific and comprehensive theoretical model of bankruptcy was proposed by Hudson [1986]. He argued that the involuntary bankruptcy of a firm results from three crises: liquidity crisis, profitability crisis and net worth crisis. Koh [1987] developed six relevant financial ratios based on this theory. A liquidity crisis (short-term and long-term) refers to the situation where a firm is unable to discharge its short-term and long-term liabilities respectively. A short-term liquidity crisis happens when the ratios of a firm's quick assets to current liabilities is less than 1. In contrast to a short-term liquidity crisis, a long-term liquidity crisis arises when the difference between a firm's market value of equity to total assets and total liabilities to total assets is negative. Therefore, the ratios for liquidity crisis are denoted as

$$(a) \quad QA < CL \quad QA/CL < 1 \quad \text{where } QA = \text{quick assets, } CL = \text{current liabilities}$$

$$(b) \quad MV < TL \quad MV - TL < 0 \quad \text{where } MV = \text{market value of equity, } TL = \text{total liabilities}$$

Standardising by total assets (TA), it becomes:

$$\frac{MV}{TA} - \frac{TL}{TA} < 0$$

A profitability crisis occurs when a firm is unable to generate sufficient earnings. As in the case of liquidity, there are two types of profitability crisis: short-term and long-term. A short-term profitability crisis exists when the earnings generated by a firm are insufficient to cover its interest payments. In other words, a short-term profitability crisis arises when the ratios of a firm's earnings before interest and tax to interest payment is less than 1. On the other hand, a long-term profitability crisis happens when the net income generated by a firm is insufficient to provide a return on investment. It can be represented as a situation when a



firm's net income to total assets is less than the return required by investors. Therefore, ratios for profitability crisis are denoted as

$$(c) \text{ EBIT} < \text{IP} \quad \text{EBIT/IP} < 1 \quad \text{where EBIT= earnings before interest and tax}$$

$$\text{IP} = \text{interest payments}$$

$$(d) \text{ NI/TA} < K \quad \text{where NI= net worth, TA= total assets,}$$

$$K = \text{required rate of return}$$

Net worth crisis refers to the difference between total assets and total liabilities or shareholders' equity, which comprises mainly common stock and retained earnings. It exists when shareholders' equity is negative. In other words, a net worth crisis means a situation where the total of common stock plus retained earnings is negative. That is

$$(e) \text{ CS} + \text{RE} < 0 \quad \text{where CS = common stock, RE = retained earnings}$$

If it is standardising by total assets (TA), the condition becomes

$$\frac{\text{CS}}{\text{TA}} + \frac{\text{RE}}{\text{TA}} < 0$$

Combining equations (a) to (e), the justification for selecting the independent variables [Koh, 1987] based on the Hudson's financial crisis model [1986] can then be defined in terms of the following financial ratios: (1) quick assets to current liabilities, (2) market value of equity to total assets, (3) total liability to total assets, (4) interest payments to earnings before interest and tax, (5) net income to total assets, and (6) retained earnings to total assets.

Instead of analysing the reasons for bankruptcy, Laitinen [1991] developed his bankruptcy models by discriminating between different failure processes. He claimed that different financial ratios should be applied to correspond to different failure processes, and argued that the previous failure prediction model based on an assumption of a common uniform process might result in inaccuracy, since the optimal failure prediction model for each process may be different according to different financial ratios and different weights for these ratios. The predictive capabilities of a failure prediction model would be determined by the frequency of each distinctive failure process in the sample. The aim of Laitinen's

paper was to select financial ratios on the basis of a theoretical model to show the important dimensions or factors which affect the financial ratios. The first type failure is called chronic failure. A firm is said to be chronically insolvent if it becomes increasingly unable to meet its financial obligations over two or more accounting periods. In such a case, the failure signal has been revealed several years prior to bankruptcy, and many financial ratios, such as return on investment, cash flow to net sales, total debt ratio to assets ratio and current ratio, has shown increasingly deterioration. The second type of failure refers to poor revenue financing. This failure is primarily due to unfavourable profitability and slow accumulation of revenue. There is no significant difference in the debt to total assets ratio and the current ratio between failed and nonfailed firms several years prior to failure. But failure can be revealed from adverse inventory turnover and assets turnover or the ratio of cash flow to net sales ratio in this firm. The failure of these kinds of firms can be predicted with a high accuracy in the second year before failure. However, the predictive power becomes rather unreliable when earlier data are used. The third type of failure is most difficult to predict owing to the abrupt reversal of the financial situation just prior to bankruptcy. A firm like this is seen to be acute insolvency. There are no statistically significant differences in the financial ratios between the failed and nonfailed firms until the first year before failure. The only sign of failure before this year may be rather poor revenue financing measured by cash flow to net sales ratio. This is different from simple cash famine. In this case a firm not only has insufficient cash in the present or short term (under one year's time) to meet its financial obligations as they fall due, and suffers from inadequate power in the collection of receivables, but also the firm appears to have nil or insufficient internal financing mechanisms available to cover the latest historical losses. In other words, in the last year for this kind of firm, almost all financial ratios deteriorate dramatically and thus can not be predicted earlier.

In Laitinen's paper [1991] he developed a model to depict what are the minimum basic dimensions or factors required to define these above processes. There are several implications in his theoretical analysis

- (1) The profitability of a firm is a very important factor that affects all the financial ratios considered in his context. This profitability dimension is measured by the return on investment.

- (2) From the perspective of the failure process, the rate of growth tends to be a relevant dimension. The rate of growth in total assets can represent this dimension.
- (3) The interactions of the profitability and growth determines the cash flow to net sales ratio.
- (4) The rate of revenue accumulation that affects the sufficiency of revenue finance is also an essential element for measuring the failure risk. The net sales to total assets ratio can be used to refer to the dimension of revenue accumulation.
- (5) The loan-taking intensiveness of a firm, which refers to the propensity of the firm to use debt capital to finance its expenditure, is an additional important dimension, which may be independent of the other four dimensions. It can be presented by the debt to assets ratio.
- (6) The harmony coefficient of debt financing, which depicts the harmony between the debt and asset structures of a firm, is measured by current ratio.

These six theoretically identified factors were also supported by empirical studies developed by Pinches, Mingo and Caruthers [1973]. For example, the return investment pattern in the latter study may refer to the profitability factor in Laitinen's research, capital intensiveness to the factor with the same label, financial leverage to the loan-taking intensiveness factor, and short-term liquidity to the harmony of the debt financing factor.

In Laitinen's work [1991], 40 randomly selected failed companies and their nonfailed mates consisting of the aforementioned three types of failure process were identified by factor analysis. Two important findings came out of his study. First, the predictive power of the discriminant model did not significantly improve when the number of financial ratios was increased to 20. Thus he concluded that the six ratios used may represent all the relevant dimensions underlying financial distress. Secondly, the prediction accuracy was dependent upon the frequencies of alternative failure processes in the sample. In this study the proportion of "chronic failure", "revenue financing failure firms", and "acute failure firms" in the sample were 32.5, 27.5, and 40.0 percent respectively. The predictive ability was thus dominated by the large proportion of acute failure firms in the entire sample and deteriorated considerably in the second year before failure. Hence, the characteristics of the sample played a key role in failure prediction. Additionally, the empirical results also

showed there is a connection between the size, business branch, and the type of failure process. For example small businesses may include more "acute firms" than others, which leads to different results in failure prediction. The findings in Laitinen [1991] provide us with a good insight and perspective for understanding and selecting the financial ratios in bankruptcy prediction.

### **8.2.3 The Rationale for Choosing the Financial Ratios in this Study**

The selection of financial ratios in our empirical study is based primarily on the theoretical considerations we have just presented in the preceding section. Although a large number of financial ratios can be computed and used to describe the underlying characteristics or attributes of a firm, they may cause the duplication of information resulting from highly correlated financial ratios or the problem of multicollinearity. Our model, using the dimensions with high relevance, thus the much smaller ratios set, provides justification for choosing independent variables, and avoid the selection bias through arbitrary and subjective means.

The Hudson [1986] and Laitinen [1991] theoretical models are combined to obtain the relevant 10 financial ratios. They are: (1) quick assets to current liabilities ratios, (2) market value of equity to total assets, (3) total liability to total assets, (4) the ratio of interest payment to earnings before interest and tax, (5) net income to total assets, (6) retained earnings to total assets, (7) cash flow to net sales ratio, (8) net sales to total assets, (9) debt to assets, (10) current assets to current liabilities. In effect these ten predictor variables developed from two theoretical models can be regarded as representing both the vertical (failure process) and horizontal (failure cause) views of business failure prediction. According to the grouping by Chen and Shimerda [1981], who have analysed the main studies and tabulated the frequency of individual ratios and the main factors involved, these ten financial ratios can be classified as Table 8.2.2.

We note that none of the variables based on the theoretical model is classified into the inventory turnover and receivable turnover group. To include these two groups in the model, two other financial ratios are selected on the basis of popularity in the literature.

The sales to working capital ratio represents the inventory turnover group, while the quick assets to sales ratio refers to the receivable turnover.

The above 12 financial ratios will be used as independent variables in this empirical study. We do not continue to employ factor analysis because the financial dimensions derived from factor analysis, and the financial dimensions that best discriminate among the group, are not necessarily the same, as we demonstrated in Chapter Five. Likewise, we do not use the stepwise procedure because this would favour the MDA, since it formulates an optimal discriminant function that also maximises the 'distance' between the group. Therefore, to have a fair basis for comparison, the same 12 financial ratios will be applied for the four alternative techniques.

**Table 8.2.1 The Grouping of Financial Ratios Selected in Empirical Study**

Grouping	Ratios	Name
1. return on investment (profitability)	(5) net income to total assets	R1
2. capital turnover	(2) market value of equity to total assets	R2
	(8) net sales to total assets	R3
3. financial leverage	(3) total liability to total assets	R4
	(6) retained earnings to total assets	R5
4. liquidity	(1) quick assets to current liabilities ratios	R6
	(9) debt to assets	R7
	(10) current assets to current liabilities.	R8
5. cash position	(7) cash flow to net sales ratio	R9
6. inventory turnover	(11) sales to working capital ratio	R10*
7. receivable turnover	(12) quick assets to sales ratio	R11*
8. other ratios	(4) the ratio of interest payment to earnings before interest and tax	R12

\*denotes the variable is not selected based on the theoretical model

### **8.3 Problems with the Choice-Based Sampling Design**

With most existing estimation methods, an exogenous random sampling process is an implicit assumption. That is, an observation is randomly drawn and the dependent and independent variables are observed. For example, in the bankruptcy prediction problem,

the exogenous sampling involving a sequence of firms is drawn and their behaviour is observed. In contrast, in a choice-based sampling process, a sequence of bankrupt firms (chosen alternatives) are drawn and the characteristics of the firms selecting those alternatives are observed. From another point of view, a choice-based sample occurs when the probability of an observation entering the sample depends on the values (attributes or group) of the dependent variables. More specifically, in most published models, the samples were drawn on the basis of knowledge of dependent variables (i.e., bankruptcy or nonbankruptcy) instead of decision makers themselves.

As Manski and Lerman stated [1977], the reason why the choice-based sample process was employed in most studies was that data collection costs for such a process are often considerably smaller than for exogenous sampling. However, estimating models based on such nonrandom samples can result in biased parameter and probability estimates if appropriate adjustments are not used [Zmijewski, 1984]. The observed result of this bias is that a dependent variable group having a sample probability larger than the population probability is oversampled, with the oversampled group having understated classification and prediction error rates.

Previous attempts to "model" bankruptcy prediction often used relatively small nonbankrupt samples with a matched-paired choice-based sampling design. Under such a design, the sample was based mainly on the attributes of the bankrupt firms. Thus the matched-paired design produces estimated probabilities that are upward-biased because the bankruptcy event being predicted is relative rare in real world.

However, the matched-pairs design has its appeal. Beaver [1966] recommended that the matched-design should be applied to provide a "control" over factors that might otherwise obscure the relationship between ratios and failure. In Altman's well known study [1968], he investigated these problems by constructing two samples. One was of pairs matched by assets, the other was unmatched. The results indicated that the matched sample offered 96% classification accuracy for one year prior to bankruptcy, while the unmatched sample provided only 79% classification accuracy. The advantage of the matching approach seemed virtually unquestionable. Many subsequent researchers including Lev [1974], Izan [1984], Zavgren [1985] and Platt and Platt [1990] adopted this view. Jones [1987] stated that if the nonbankrupt firms were drawn at random, substantial differences between two

groups may arise in terms of industry and size. The model attempting to discriminate between failing and nonfailing firms may be transformed to distinguish between large and small firms or between different industries. Nevertheless, as pointed out by Keasey and Watson [1991], most studies dealt with sample selection by matching nonfailed firms to the failed firms by industry and size, which overcomes the issue of how to define the size of the nonfailed sample, this solution also rules out size and industry as prediction variables. But this is not the main shortcoming of matching sampling. The most detrimental drawback proven by Manski and Lerman [1977] and Manski and Mcfadden [1981] was that statistical classification models which violate the random sampling and ignore the choice-based sampling procedure lead to asymptotically bias of both the parameters and probability estimates. Consequently, very large random samples are needed to obtain information on the rare occurrences of bankruptcy if the effects of choice-based sample bias are to be minimised. Bankrupt firms were obviously oversampled in previous studies because their sample base rate (0.5) was much larger than their population probabilities. The results of this research were understatement of classification and prediction error rates in the oversampled group. In other words, by assuming equi-probability, the matched sample design "magnifies" the effects of the choice-based sample bias.

The question which arises here is how we can, on the one hand, take advantage of matched sampling design; and on the other hand, avoid the disadvantage of its bias. Some adjusted techniques may be suggested in order to achieve this goal.

### **8.3.1 Weighted Exogenous Sample Maximum Likelihood (WESML)**

In order to correct the oversampling bias, incorporating an adjustment procedure such as the weighted exogenous sample maximum likelihood (WESML) approach into the estimation technique seems to offer a better solution [Manski and Lerman, 1977]; [Manski and Mcfadden, 1981]. The essence of WESML is modifying the sampling maximum likelihood estimator by weighing each observation's contribution to the log-likelihood. That is, WESML weights the estimation function according to the proportion of bankrupt firms in the sample  $\alpha_s$  and in the population  $\alpha_p$ . For the Logit model, without adjustment

procedure, the parameters are estimated by maximising the log-likelihood function (L) as shown below

$$L = \sum \ln P(\beta) + \sum \ln[1-P(\beta)] \quad (8.3.1)$$

where

$$P(b) = F(Z) = \frac{1}{1+e^{-Z}}$$

$$Z = \mathbf{b}'\mathbf{X} + e$$

The equation above assumes that the sample proportions of the two groups equal those of the population. With the WESML procedure, the differences between sample and population proportions can be adjusted by maximising the weighted log-likelihood function

$$L' = (\alpha_p/\alpha_s) \sum \ln P(\beta) + [(1-\alpha_p)/(1-\alpha_s)] \sum \ln[1-P(\beta)] \quad (8.3.2)$$

where

$\alpha_p$  = the proportion of bankrupt firms in the population

$\alpha_s$  = the proportion of bankrupt firms in the sample

Comparing the two equations for L and L', the further the sample proportion away from the population, the more distorted the model results will be. When  $\alpha_s$  approaches  $\alpha_p$ , L' approaches to L. As the  $\alpha_s$  equals  $\alpha_p$ , there is no more bias in the estimation process.

By using the WESML procedure, the matched-pair sample design can be used to control for industrial and size effect. On the one hand, it deletes the factors which are unrelated to the phenomenon investigated. On the other hand, it eliminates the biased coefficient estimates which encountered with this matching choice-based sampling design.

The WESML was first applied by Zmijewski [1984] in the construction of bankruptcy models. Zmijewski conceptually and empirically examined choice-based sampling biases arising from oversampling bankrupt firms in bankruptcy prediction studies. In order to investigate the impacts of weighted and unweighted processes, a series of samples consisting of 40 bankrupt and from 40 to 800 nonbankrupt firms were employed for Probit analysis on six different data sets. Using the unweighted Probit, a higher accuracy rate of 97.2% was attained in classifying bankrupt firms and a lower rate of 92.5% in classifying nonbankrupt firms in the holdout sample within the 40:40 sample proportion. As the sample proportion moved closer to the population proportion, the accuracy rates with



regard to the failing firms decreased, and the accuracy rates with regard to the nonfailing firms increased. At the sample proportion was at 40:800, bankrupt firms were correctly classified only 71% against the 99.5% correctly classified in nonbankrupt firms. According to the Pearson correlation coefficient, Zmijewski found that there was a significant relationship between sample proportions and prediction accuracy. Clearly, the evidence showed that when the sample proportions of bankrupt and nonbankrupt firms do not reflect the prior probability of these two groups in the population, an unweighted Probit analysis overstates the classification accuracy for bankrupt firms and understates the accuracy for nonbankrupt firms. However, these biases decrease as the sample selection probability approaches the population probability. Then, Zmijewski [1984] applied the WESML procedure to the same range of above sample proportions. He fixed 0.847% as the prior probability of bankruptcy, which was considerably smaller than the smallest sample proportion 4.76% used in the test of unweighted Probit. In the 40:40 case, the results indicated that the model provided 54% and 99.8% classification *accuracy in bankrupt and nonbankrupt firms* respectively. In the 40:800 sample, the accuracy was 43.9% and 100% respectively. Additionally, the Pearson correlation coefficient has indicated statistically insignificant outcomes in the relationship between sample proportion and prediction accuracy. The evidences confirmed the great extent of elimination of choice-based sampling biases by using the WESML procedure. These results were consistent with the studies of Manski and Lerman [1977].

Dopuch, Holthausen, and Leftwich [1987] subsequently constructed models to predict audit qualifications for going-concern status. A Probit analysis with a WESML technique to eliminate the bias in the coefficient estimates resulted from choice-based sampling. The sample contained 218 qualified opinions and 346 clean opinions for the fiscal year 1973-1975. The proportion of qualified opinion in the population  $\alpha_p$  was estimated by the Markov transition matrix depending on the total number of firms with data available in the population per year, the number of first-time qualifications per year and the transition probability. The range of  $\alpha_p$  value was calculated from 0.02622 to 0.1078. The results for five value 0.02622, 0.03859, 0.04689, 0.06982 and 0.10718 were reported. For comparison purposes, the Probit analysis for the five  $\alpha_p$  WESML models, the unweighted Probit analysis and the OLS (ordinary least squares) regression were predicted. The cutoff

points were selected to minimise the total misclassification cost for alternative costs of Type I to Type II errors, namely: 1:1, 5:1, 10:1 and 20:1. The evidence indicated that the performance of the models is not sensitive to the particular choice of  $\alpha_p$  within the above range. While the estimated probabilities and parameters inference of qualified and clean firms are quite different, for prediction purposes, the unweighted Probit and OLS regression were almost equivalent to the WESML results.

Although, according to the above two studies, using the adjusted estimation technique may not improve overall classification and prediction error rates, it will provide unbiased parameter and probability estimates allowing users both to assess the effect of individual variables and to choose optimal cutoff probabilities based on individual loss function [Zmijewski, 1984]. Based on the above discussion, this thesis will further explore this choice-based sampling bias problems among MDA, Logit procedure and two artificial neural network algorithms.

### 8.3.2 Proposed Solution to Determine Optimal Cutoff Point

Although WESML procedure can eliminate the choice-based bias, it usually involves a complicated computation when maximising the adjusted likelihood function. More importantly, in some methods there are no such likelihood functions to be maximised.

In order to overcome this problem, a proposed solution is presented. This solution provides the change of optimal cutoff point instead of changing the likelihood function.

The proposed approach is indicated in the following:

Under the objective of minimising the total misclassification cost, the mathematical equation can be expressed

$$\text{Min } E(c) = C_I \alpha_p \int_{\tau_1} f(I/B) dI + C_{II} (1 - \alpha_p) \int_{\tau_2} f(I/N) dI \quad (8.3.3)$$

where

B = bankrupt firms

N = nonbankrupt firms

$\alpha_p$  = the prior probability of bankruptcy in population

I = the index value in discriminating model

I\* = the optimal cutoff point

$t_1$  = the region in the domain of  $I$  within which we assign  $I$  belongs to nonbankruptcy  
 $t_2$  = the region in the domain of  $I$  within which we assign  $I$  belongs to bankruptcy  
 $C_I$  = the cost of misclassifying a bankrupt firm as a nonbankrupt firm,  
           i.e. misclassification cost of Type I error  
 $C_{II}$  = the cost of misclassifying a nonbankrupt firm as a bankrupt firm,  
           i.e. misclassification cost of Type II error  
 $f(I/B)$  = the probability density function of  $I$  for bankrupt firms  
 $f(I/N)$  = the probability density function of  $I$  for nonbankrupt firms

By using a weighted estimation concept, the  $\alpha_s$  will be embedded in the above equation. That is, the equation is divided by  $\alpha_s$  and  $(1-\alpha_s)$ , which has the same form as the weighted log-likelihood function in equation (8.3.2)

$$\text{Min AC} = C_I (\alpha_p/\alpha_s) \int_{\tau_1} f(I/B) dI + C_{II} [(1-\alpha_p)/(1-\alpha_s)] \int_{\tau_2} f(I/N) dI \quad (8.3.4)$$

For ease of reference the equation is called adjusted total cost AC. The addition of  $\alpha_s$  into the denominators is used to adjust the impact of prior probability. The further the  $\alpha_s$  is away from the  $\alpha_p$ , the stronger is the impact of the prior probability on the adjusted total cost. When the  $\alpha_s$  is equal to  $\alpha_p$ , the minimisation of adjusted total cost is equivalent to the minimisation of total cost since the *impact of prior probability is diluted by the identical* proportion of these two groups in the sample.

The optimal cutoff value  $I^*$  is found by differentiating AC with respect to  $I$  to minimise the adjusted total cost [Hsieh, 1993]; [Steele, 1995].

Since  $\tau_1$  and  $\tau_2$  are independent and cover the whole domain, we have

$$\int_{\tau_1} f(I/B) dI = 1 - \int_{\tau_2} f(I/B) dI$$

The adjusted total cost (AC) can then be written as :

$$\begin{aligned}
 AC &= \int_{\tau_1} C_I (\alpha_p/\alpha_s) f(I/B) dI + \int_{\tau_2} C_{II} [(1-\alpha_p)/(1-\alpha_s)] f(I/N) dI \\
 &= C_I [1 - \int_{\tau_2} f(I/B) dI] (\alpha_p/\alpha_s) + \int_{\tau_2} C_{II} f(I/N) dI [(1-\alpha_p)/(1-\alpha_s)] \\
 &= C_I (\alpha_p/\alpha_s) + \int_{\tau_2} (-C_I)^* (\alpha_p/\alpha_s)^* f(I/B) dI \\
 &\quad + \int_{\tau_2} (C_{II})^* [(1-\alpha_p)/(1-\alpha_s)]^* f(I/N) dI \\
 &= C_I (\alpha_p/\alpha_s) + \int_{\tau_2} [(-C_I)^* (\alpha_p/\alpha_s)^* f(I/B) \\
 &\quad + (C_{II})^* [(1-\alpha_p)/(1-\alpha_s)]^* f(I/N)] dI \quad (8.3.5)
 \end{aligned}$$

To obtain the optimal point  $I^*$ , it is necessary to choose  $\tau_1$  and  $\tau_2$ , such that AC is minimised, since  $C_I(\alpha_p/\alpha_s) > 0$  for a given AC, if the integrand of (8.3.5) is negative, AC will be decreased by assigning this  $I$  to  $\tau_1$  and vice versa.

Consequently, the decision rule to select the optimal cutoff point can be written as

$$\begin{aligned} (-C_I) * (\alpha_p/\alpha_s) * f(I/B) + (C_{II}) * [(1-\alpha_p)/(1-\alpha_s)] * f(I/N) &= 0 \\ (C_{II}) * [(1-\alpha_p)/(1-\alpha_s)] * f(I/N) &= (C_I) * (\alpha_p/\alpha_s) * f(I/B) \\ f(I/N) / f(I/B) &= (C_I/C_{II}) * (\alpha_p/\alpha_s) / [(1-\alpha_p)/(1-\alpha_s)] \\ &= (C_I/C_{II}) * [(\alpha_p/(1-\alpha_p)) * [(1-\alpha_s)/\alpha_s]] \end{aligned}$$

where

$\alpha_p$  = the proportion of bankruptcy to nonbankruptcy in the population

$\alpha_s$  = the proportion of bankruptcy to nonbankruptcy in the sample

Put another way, the marginal condition for optimal cutoff point  $I^*$  is given by

$$\frac{f(I^*/B)}{f(I^*/N)} = \frac{C_{II}}{C_I} \times \frac{1-\alpha_p}{\alpha_p} \times \frac{\alpha_s}{1-\alpha_s} \quad (8.3.6)$$

At the optimum, the cutoff point  $I^*$  depends on the estimates of three parameters: (1) the ratio of judgements about the costs of misclassification; (2) the ratio of prior probability of nonfailing and failing firms (3) the ratio of the proportion of failing and nonfailing firms in the sample.

According to the above analysis, the use of weighted adjustment procedure in MDA, Logit and neural network methods can be achieved by changing the selection of the optimal cutoff point. For instance, if the ratios of judgements about misclassification costs, prior probabilities and base rate are denoted as  $K$

$$K = \frac{C_{II}}{C_I} \times \frac{1-\alpha_p}{\alpha_p} \times \frac{\alpha_s}{1-\alpha_s}$$

Then the optimal cutoff point occurs when

$$f(I^*/B) = K f(I^*/N)$$

The optimal condition may be displayed Graphically as in Figure 8.3.1

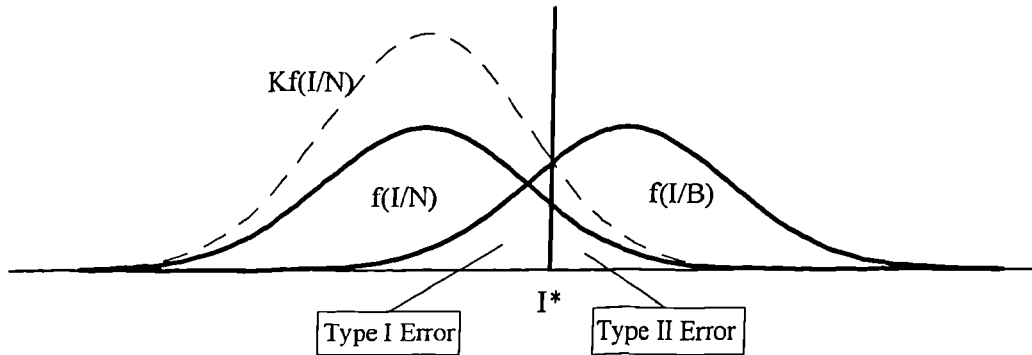


Figure 8.3.1 Hypothetical Frequency Distribution of Index I

Under the assumption of equal misclassification cost, equal prior probability and one to one matched pair choice-based sampling procedure, the value of K is 1. That means,

$$\frac{f(I^*/B)}{f(I^*/N)} = 1$$

The optimal cutoff point  $I^*$  from the decision is when the ordinates of posterior (or conditional probability) are equated. In terms of MDA's Z score model, this is equivalent to setting the optimal cutoff point to a midpoint of  $\bar{Z}_1$  (the average value of Z score (in population 1) and  $\bar{Z}_2$  (the average value of Z score in population 2), when the variances of two groups are identical. As to the Logit process and two neural network algorithms with sigmoid function (logistic c.d.f.) in the output layer, this is equivalent to setting the optimal cutoff point as a benchmark conditional probability value 0.5 in the Logit method or a 0.5 predicted value in the output layer of neural networks for binary choice [Tam and Kiang, 1990].

When the K value does not equal 1, the optimal cutoff point  $I^*$  can be estimated in the following way

1. If the distributions of index I both for bankrupt and nonbankrupt groups are normal distributions with mean  $\mu_1$ ,  $\sigma_1$  and  $\mu_2$ ,  $\sigma_2$  respectively, then

(1) Firstly, a logarithmic transformation is performed for equation (8.3.6), thus the equation become

$$\ln f(I^*/B) - \ln f(I^*/N) = \ln\{(\alpha_s/\alpha_p)[(1-\alpha_p)/(1-\alpha_s)] (C_{II}/C_I)\} \quad (8.3.7)$$

$$\begin{aligned} \text{and } \ln f(I^*/B) &= \ln[1/\sigma_1 \sqrt{2\pi} \exp[-1/2(I^* - \mu_1/\sigma_1)^2]^2 \\ &= -\ln \sigma_1 - \ln \sqrt{2\pi} - 1/2[I^* - \mu_1/\sigma_1]^2 \end{aligned}$$

Likewise,

$$\begin{aligned} \ln f(I^*/B) &= \ln[1/\sigma_2 \sqrt{2\pi} \exp[-1/2(I^* - \mu_2/\sigma_2)^2]^2 \\ &= -\ln \sigma_2 - \ln \sqrt{2\pi} - 1/2[I^* - \mu_2/\sigma_2]^2 \end{aligned}$$

Thus,

$$\ln f(I^*/B) - \ln f(I^*/N) = \ln(\sigma_2/\sigma_1) + (I^* - \mu_2)^2/2\sigma_2^2 - (I^* - \mu_1)^2/2\sigma_1^2 \quad (8.3.8)$$

(2) Combine equation (8.3.7) and (8.3.8), we obtain

$$\begin{aligned} \ln f(I^*/B) - \ln f(I^*/N) \\ = \ln(\sigma_1/\sigma_2) + \ln(\alpha_s/\alpha_p) + \ln[(1-\alpha_p)/(1-\alpha_s)] + \ln(C_{II}/C_I) \end{aligned} \quad (8.3.9)$$

2. If the distributions of index I both for bankrupt and nonbankrupt groups are not normal distributions, we can use a numerical method such as Newton's iteration method to solve it.

To find a solution to an equation of the form  $f(I) = 0$ , the Newton's method starts at  $I_0$ , then uses knowledge of derivative  $f'$  to take a sequence of steps toward a solution. Each new point  $I_n$  that it tries is found from the previous point  $I_{n-1}$  by the formula  $I_n = I_{n-1} - f(I_{n-1})/f'(I_{n-1})$ . For instance, when  $f(I/B)$  and  $f(I/N)$  are logistic functions, then

$$f(I) = \frac{\frac{1}{b} \exp(-(\frac{I-a}{b}))}{[1 + \exp(-(\frac{I-a}{b}))]^2}$$

I is a variable from logistic distribution

a any value,  $b > 0$

$$F(I) = \frac{1}{1 + e^{-I}}$$

F is the cumulative distribution function of I

$$E(I) = a$$

E(I) is the population means m

$$\text{Var}(I) = \frac{b^2 \pi^2}{3}$$

Var(I) is the population variance s

Then, the marginal condition for optimum is

$$\frac{f(I^*/B)}{f(I^*/N)} = \frac{\frac{1}{b_1} \exp(-(\frac{I^*-a_1}{b_1}))}{(1+\exp(-\frac{I^*-a_1}{b_1}))^2} \div \frac{\frac{1}{b_2} \exp(-(\frac{I^*-a_2}{b_2}))}{(1+\exp(-\frac{I^*-a_2}{b_2}))^2} = K \quad (8.3.10)$$

$$\text{where } a_i = \mu_i \quad b_i = (\sigma_{ii}/\pi)\sqrt{3}$$

Given the  $k$  value and other necessary parameters, the root  $I^*$  of the above equation will be found through numerical methods.

Hence the optimal cutoff point  $I^*$  can be numerically derived by solving either equation (8.3.9) or (8.3.10) as long as all parameters  $\mu_1, \mu_2, \sigma_1, \sigma_2, \alpha_s, \alpha_p, C_{II}, C_I$  are obtained. The parameters  $\mu_1, \mu_2, \sigma_1, \sigma_2$  are usually empirically estimated from the sample of bankrupt and nonbankrupt firms. The  $\alpha_s, \alpha_p, C_{II}, C_I$  might depend upon the historical data or subjective judgement.

For ease of reference, the above equations (8.3.9) and (8.3.10) are called the weighted cutoff point (WCOP) procedure. By applying this decision model in bankruptcy prediction, the validity of the classification accuracy is enhanced because choice-based bias is minimised or eliminated. This process requires neither any assumptions about the distribution of independent variables nor the restriction of their variance matrices. Thus the results can be applied to conventional statistical techniques as well as to the artificial neural network approaches.

To investigate the influences of the parameters  $\alpha_p, \alpha_s, C_I, C_{II}$ , on the change of the optimal cutoff point, a simpler way is to view the following condition as developed previously:

$$\frac{f(I^*/B)}{f(I^*/N)} = \frac{C_{II}}{C_I} \times \frac{1 - \alpha_p}{\alpha_p} \times \frac{\alpha_s}{1 - \alpha_s} = K$$

As the value varies according to the size of the prior probability in population ( $\alpha_p$ ), the base rate in sample ( $\alpha_s$ ), or changes in the misclassification costs ( $C_I$  and  $C_{II}$ ), so will the optimal cutoff point  $I^*$  change. Indeed it is apparent that the classification accuracy of a discriminating technique depends entirely on the value  $K$  [Steele, 1995].

The determination of the optimal cutoff point  $I^*$  can be explained in Figure 8.3.1. The left distribution represents the conditional probabilities of failure for the healthy group, and the right one is the distribution of the conditional probabilities of failure for the bankrupt group. We can see from Figure 8.3.1 that the lower the prior probability of bankruptcy in the population  $\alpha_p$ , the higher the base rate in sample  $\alpha_s$ , and the smaller the ratio of  $C_I$  to  $C_{II}$ , the larger  $K$  value will be, and thus the more degree in right shift of the optimum cutoff point will be, and the greater the Type I error will be obtained.

The proposed adjustment procedure WCOP avoids the complicated calculation originally produced by weighted maximum likelihood function, and provides a feasible solution especially for the case in which prior probabilities and base rate cannot be embedded into the neural network model. More importantly, since the probability density function of the independent variables is not limited to a specific distribution in the above process, this result can be adopted in different approaches.

#### 8.4 Problems with the Unequal Misclassification Costs

In the light of the above discussion, the optimal points criterion is determined by three factors. In addition to the base rate in the sample, and the prior probability in the population, two kinds of misclassification costs are also essential. When Type I and Type II error costs are identical, minimising the total error probability is the same as minimising the total error cost. However, when the loss functions of the errors are asymmetrical, minimising the total error probability is not the same as minimising the total error cost; and the outcome will depend heavily on the cost ratio of Type I and Type II errors. Zmijewski [1983] analysed this impact on the predictive ability of the bankruptcy prediction model. He noted that when the Type I error cost is twice that of a Type II error cost, the more accurate predictions are Beaver's [1966] and Blum's [1974]. When the ratio is changed to 20 to 1, the most accurate model is Ohlson's [1980]. Merely considering the prior probability and assuming equal costs of Type I error and Type II error in most prior studies often resulted in emphasising the correct prediction of nonbankrupt firms at the expense of bankrupt firms, since nonbankruptcy occurs much frequently than bankruptcy. In other



words, there is a stronger stress on reducing Type II errors at the expense of Type I error. Type I error cost is defined as the cost of misclassifying a bankrupt firm as a nonbankrupt firm. It usually involves the cost of holding a long position in the equity securities of failing firms. The investors lose the entire investment, which can be substantial. The Type II error cost is defined as the cost of misclassifying a nonbankrupt firm as a bankrupt firm. It can be estimated by the opportunity cost of losing the dividend income and capital gain that may be obtained otherwise.

Altman [1980] has empirically estimated these two kinds of misclassification cost by sending questionnaires to over 400 bank representatives. He approximated the cost of misclassifying a bankrupt firm as nonbankrupt through the use of the cost to the commercial banks of accepting that default. The opportunity cost of rejecting a loan that would have resulted in a successful payoff was employed to approximate the cost of misclassifying a nonbankrupt firm as a bankrupt firm. The loss rate on the defaulted loans was approximately 62 percent of the loan principal, while the opportunity cost of the decision not to lend to an account that would have repaid successfully was estimated to be 2 percent. That is, the cost of misclassifying a bankrupt firm as a nonbankrupt firm is almost 31 times more costly than the cost of misclassifying a nonbankrupt firm as bankrupt. Hsieh [1993] also performed an empirical study of the  $C_I/C_{II}$  ratio. A sample of 43 actually bankrupt firms and 43 healthy firms matched for beta and size from 1968-1976 were examined. The Type I error cost was estimated as the return difference between the return which could have been earned from correctly selling short failing securities and the return actually earned from incorrectly holding a long position in these failing firms. Type II error cost is, then, the difference in return between the correct action of holding a long position in the equity securities of healthy firms and the incorrect action of holding a short position in these securities. The estimated result of the  $C_I/C_{II}$  ratio was about 3.2419, calculated by applying the data of stock prices, dividends, returns on the market portfolio, transaction costs and a broker's interest charge.

The above cases stand for the positions of investors or banks. For auditors in accountancy profession, the situation is on the contrary. Although embarrassment is acute when a client company fails soon after receiving a clean audit opinion on its financial statements, the more serious pressure they face is the litigation of wrongly declaring bankruptcy and the

lose of business. As St Pierre and Anderson [1984] found no legal case in which an auditor has been found negligent for excessive conservatism. Thus the possible smallest Type II error is desirable for the auditors.

Our research explores these asymmetric loss functions from the banks point of view. While all prediction errors are undesirable (lose capital or lose business), it is generally accepted that the incorrect prediction of a bankrupt firm as nonbankrupt is the costlier error. Thus, ignoring the unequal misclassification is inappropriate when the optimal cutoff point (probability) is determined. In Figure 8.4.1, without considering the error costs,  $I_0$  is the optimal cutoff point which minimises the total error probability. However, when Type I and Type II error costs are taken into account, the optimal decision model shifts the cutoff point to the left to  $I^*$  to reduce the Type I error probability since the Type I error cost is much more costly than the Type II error cost. In particular, the higher the misclassification cost of a Type I error relative to that of a Type II error, the more the Type I error should be minimised in choosing the optimal cutoff point. At the extreme, when the misclassification cost of Type I error is infinity, the optimal cutoff point must be set such that no Type I error occurs.

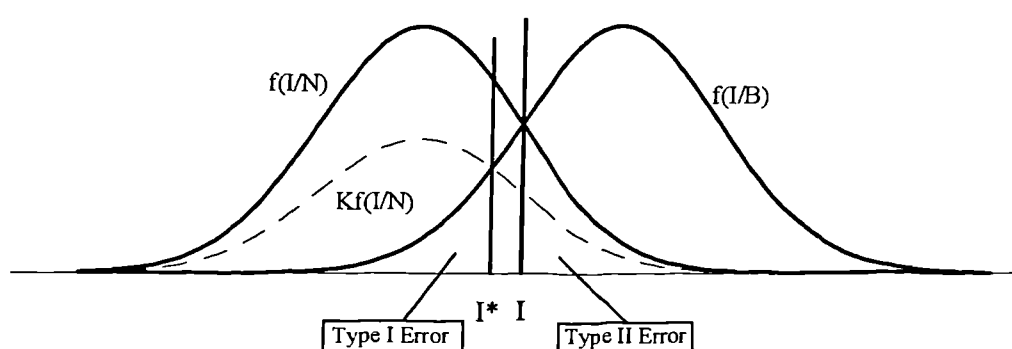


Figure 8.4.1 Change of Optimal Cutoff Point to the Ratio of  $C_{II}$  to  $C_I$

A few bankruptcy prediction model studies have tried to incorporate Type I and Type II error costs into the classification accuracy test. As mentioned above, Dopuch, Holthausen and Leftwich [1987] have considered the influences of the alternative misclassification costs in determining the optimal cutoff point for the adjusted Probit going-concern models.

However, the sensitivity of optimal cutoff points to misclassification costs of Type I and Type II errors has not been explored. Altman, Halderman and Nurayana [1977] considered accuracy only under a wide range of cost ratio specifications (i.e., Type I and Type II error cost ratio are 1:1, 10:1, 20:1, 30:1, 40:1, etc.) and failed to use a systematic method to derive an optimal cutoff point which minimises the total error costs. Thus, the impact of change of the optimal cutoff points to unequal misclassification costs on a decision model still remains unclear. Koh [1992], noting this important issue, extensively examined the sensitivity analysis of optimal cutoff points to these two type of misclassification costs in the going-concern optimal prediction context. The optimal cutoff points corresponding to values of  $C_I:C_{II}$  from 1:1 to 500:1 were investigated for the adjusted Logit model. The results revealed that the misclassification costs of Type I and Type II errors affect the optimal cutoff points in the going-concern prediction models. However, the optimal cutoff points that minimise the expected costs were rather insensitive to different relative costs. Consequently, Koh concluded that going-concern prediction models were generally applicable over a wide range of robustness of optimal points to misclassification costs. Koh's study was limited to the Logit model. When other approach is applied, the calculation involved in the maximum likelihood function could be very complicated, and even no such likelihood function might be applied. In contrast, the WCOP adjustment process (i.e., equation (8.3.9) and (8.3.10)) provides a feasible and systematic means applicable for all methods when deriving the optimal cutoff point for the classification using alternative cost ratios. We believe that the impact of misclassification costs on forecasting failure is a worthy area for further investigation. Thus, the sensitivity analysis of optimal cutoff points to misclassification costs of Type I and Type II errors for all our discriminating techniques is one of the interests of the present empirical study.

## **8.5 Problems with the Model Validation and Generalisation**

Researchers will often be concerned with how well the developed model fits and if it is statistically significant. With MDA, the percentage of the variation explaining the variance between groups can be measured by the canonical correlation techniques. The overall

significance of the discriminant model can be evaluated by the Wilk's lambda statistic, which has a chi-square distribution. For the Logit procedure, an overall statistical significance can be obtained by using the likelihood ratio test, which is also distributed as a chi-square.

The classification accuracy achieved in the model has been a particularly interesting subject to researchers. However, the error rate derived from the training sample has an optimistic bias and is called an apparent error rate. It is well known that a model will generally fit the sample from which it was derived, and it is not necessary to have success when applied to another sample. In order to eliminate this upward classification bias, the discriminating function should be used to classify the firms in the validation sample.

#### **8.5.1 Two Validation Methods**

This validation problem can be handled by either splitting the whole data into a derivation subsample and prediction subsample or by using the Lachenbruch cross-validation method [Lachenbruch, 1967]. In the first method, the investigated data can be divided into two sets. One set is called the training set and is for deriving the estimation function, and the other set is called holdout sample, is for predicting the error rate. Unfortunately, when the sample size is not enough, such a split-sample method has the effect of reducing the effective sample size. The parameter developed may be unreliable. In contrast, the Lachenbruch cross-validation procedure treats  $n-1$  out of  $n$  training observation as a training set. It determines the function parameter based on these  $n-1$  observation and then applies them to classify the one observation left out. This process will be repeated for  $n$  times and the percentage. This method has the advantage of producing an "almost unbiased", "almost sufficient", robust estimate of population error rate compared to the holdout method [Lachenbruch and Mickey, 1968]. However, the Lachenbruch approach has a time-consuming problem when the sample size is large. At the same time, the computational requirements of the technique might seem prohibitive in practice. For instance, in terms of MDA, the inverse for  $n$  matrices must be generated to estimate the holdout error rates and the full matrix would have to be inverted to estimate the population discriminant function. Lachenbruch and Mickey [1968] recognised these difficulties and

avoided the problem by using a matrix algebra result derived by Bartlett, which can generate the  $n$  matrix inverses with only one matrix inversion and so simplifies the operation. In effect, tedious computation is not the only problem, the more important drawback involved in this procedure is that when the holdout sample is independent of the training samples, this does not provide the test of external validity that a holdout sample offers. Jones [1987] stated that when a holdout sample is obtained from a later period, the Lachenbruch method can not test for both overfitting and a violation of the stationary assumption which implies the relationship between the independent variables and the dependent variables. Since large sample size is used in this thesis, the holdout sample procedure is the more appropriate.

### **8.5.2 *Ex post* Discrimination vs. *Ex ante* Prediction**

Even though the validation discrimination is successful, it does not imply that the successful generalisation ability in this model is achieved. The validation (*ex post*) discrimination is simply to warrant the significance of the independent variables in the discriminating function. If it is not successful, then no valid inference can be made about the explanatory power of the independent variables. It is thus, in fact, a necessary first step before ascribing explanatory importance to any of the independent variables. Joy and Tollefson [1975] argued that the true verification of the predictive content of the model requires validation outside the time period of the original sample. That is, if the model is estimated using data from time  $t$  to predict an event in time  $t+1$ , then data from a future period should be plugged into the model to predict whether an event will occur in the appropriate succeeding period in order to test the capacity to generalise. Otherwise, the classification does not constitute *ex ante* prediction, but rather *ex post* discrimination. The essence of Joy and Tollefson's argument is that a discriminating analysis model is only useful for prediction purpose if the relationships among the variables in the populations are stable over time. Otherwise, the model and the estimated accuracy will only be valid for the specific periods investigated. Joy and Tollefson [1975] pointed out that many researchers who used a validation sample from the original sample period mistakenly interpret their *ex post* classification results as indicators of the predictive accuracy of the model. They observed,

however, that "under the assumption of population stationarity over time *ex post* discrimination is tantamount to prediction. But the researcher must establish that stationarity exists" (p.727).

In our simulation study, all data is generated from the predefined population. This is equivalent to assuming that the basic underlying relationships and parameters are stable over time. Moreover, the purpose of the simulation study is to compare the classification power for different techniques and to identify the influence of the factors we intended to evaluate, thus there is no need to use *ex ante* prediction. With regard to the empirical study, the argument of Joy and Tollefson gives us a useful guide to the concept of inter-temporal validation.

### **8.5.3 The Impact of Different Base Rates between the Training and Testing Data**

In addition to the above two aspects, the generalisation ability could also be affected by the differences in the proportion of bankrupt firms to nonbankrupt firms between the training and validation data sets. This issue involves the question: if a classification model uses a training sample with a certain proportion of the two groups, does the model still provide good performance when the prior probability in the testing population has changed? As we showed in Chapter Four, Wilson and Sharda [1994] developed an experiment to test the effects of these proportion shifts on the predictive performance for various of training and validation set compositions. They compared the classification accuracy in the validation sample between discriminant analysis and neural networks. The three levels were 50/50, 80/20, and 90/10 proportions of nonbankruptcy to bankruptcy. By using a full two-factor design, there were nine different experimental cells. To gain reliable results, 20 different training-testing set pairs were generated via Monte Carlo resampling techniques from the original 129 firms. These 129 original firms consisted 65 bankrupt firms and 64 nonbankrupt firms, matched on industry and year from 1972-1982 obtained from Moody's Industrial Manuals. Five financial ratios as used by Altman in 1968, were applied. In general, when the training set and testing set had the identical proportions of two groups, a higher classification accuracy was achieved relative to different proportions between training and validation. The only exception occurred in the 80/20 proportion in the training

set and the 90/10 proportion in the validation set. This combination obtained 95.68% (in neural network) and 91.59% (in MDA) contrasted with the results of 91.0% (in neural network) and 89% (in MDA) by containing all the 80/20 proportion both in training and validation. Aggregately, the greater the difference between the training proportion and the testing proportion, the lower will be the prediction ability in generalisation. On the other hand, neural networks were shown to perform well in predicting both bankrupt and nonbankrupt firms in the learning phase when presented with equal numbers of examples in the two groups. This implies that a more accurate classification model will result when developed with a balanced training data. Accordingly, this study has suggested that "smoothing" the distribution of the training set, irrespective of the actual distribution, will provide a better model, since the composition of historical data necessary in the predictive model development cannot be controlled. In order to further understand the impact of the different base rates between training and validation samples on generalisation capacity, comparative analyses will be extensively investigated not only for MDA, Logit, but also for two ANN methods.

## **8.6 Summary and Conclusions**

Investigating predictive ability under comprehensive problem characteristics and identifying the favourable and unfavourable data conditions for alternative discriminating techniques can be regarded as the horizontal assessments of bankruptcy prediction. Vertical assessment are: discussing more deeply the problem associated with the selection of independent variables, the problem associated with choice-based sampling design, the problem associated with different misclassification costs, and the problem associated with model validation.

In this chapter, we have presented the difficulties and biases encountered with constructing failure prediction models in prior studies, and we have proposed solutions to cope with them. In addition to providing a means to select financial ratios as explanatory variables on the basis of bankruptcy theories instead of arbitrary and subjective judgement, the more

important contribution is to establish a systematic approach to eliminate the choice-based sampling bias by means of changing the optimal cutoff point. As we pointed out earlier, when a population's proportions of bankruptcy bear no relation to the proportion observed in the sample, failure to make the adjustment leads to biased coefficient estimates and misleadingly high classification accuracy rates for the bankrupt firms. The proposed solution also takes the unequal misclassification costs into account so that the sensitivity of the optimal cutoff points to relative misclassification costs of Type I and Type II errors can be evaluated. In neural network algorithms, there is no way to embody the above information in the model building. This is probably the main reason why previous studies of bankruptcy prediction developed on the ANNs, have not so far considered the factors of unequal prior probability and misclassification costs into the classification rule. The proposed approach is not limited to particular techniques and can be applied to any distributions of financial ratios. Moreover, it avoids complexity of reestimating the weighted maximum likelihood function, which was utilised in previous research but restricted to the Probit or Logit approaches. This decision model can enhance the validity of bankruptcy prediction models, and hence *future accounting and finance research*. The subjects to be studied and their research designs of empirical experiments will be presented in the next chapter.



## **Chapter Nine**

### **METHODOLOGY OF THE EMPIRICAL STUDY**

#### **9.1 Motivation of the Empirical Study**

The simulation study has been successfully performed on alternative techniques. It provided a variety of test conditions involving normal and non-normal populations, equal and unequal group dispersions, low and high correlations and different orientation schemes among indicators. On the other hand, we will use a real data in order to examine whether the results are consistent with those in the simulation procedure. One purpose of comparing the four classification methods on a real data set is to evaluate the methods in a situation which exhibits complexities and characteristics which are difficult or impossible to recreate in the statistically 'clean' environment of a simulation experiment. The other objectives attempt to fully assess the influences of variations in building failure prediction models which were not explored in simulated data.

In the second part of this thesis, the real financial data is first utilised with a full data set (264 companies) in order to estimate misclassification rates for those four techniques. The experiment following the simulation study is performed under the assumptions of equal prior probability and equal misclassification costs.

Secondly, in order to explore the impact of the sample size, resampling data sets with small, medium and large sample size are generated via Monte Carlo skill to evaluate its effect on the predictive abilities of the four techniques.

Thirdly, the impacts of different base rates (the proportion of bankrupt to nonbankrupt firms) on Type I error rate, Type II error rate and Overall error rate are evaluated by four methods. The key idea of this study is expressed in the question: if a researcher chooses different matched criteria in sampling design (for instance, 1:1, 1:5 and 1:10) from the same population, does there exist a significant functional relationship between classification accuracy and various sample selection probabilities? In effect, this question involves the bias problem derived from a choice-based sample design. The adjustment procedure,

weighted cutoff point (WCOP) process proposed in Chapter Eight, is applied to assess if this possible bias can be minimised.

Fourthly, breaking the assumptions of equal misclassification costs of Type I and Type II errors, a sensitivity analysis of optimal cutoff points to different ratios of two misclassification costs is studied for each of the four techniques.

Finally, we are concerned with the influence of different proportions of two groups (i.e. base rate) between training and testing samples. This concern involves the generalisation issue. When the base rate used to build the classification rule is different from that in the testing model, is there any particular technique which is less affected in this situation?

We hope that these empirical studies will provide us not only a comparison of classification power of different discriminators, and a contrast with the simulation conclusions, but will also enhance our knowledge of the sensitivity and possible bias of some factors in constructing a failure forecasting model. This knowledge is related to both traditional statistical tools and for modern neural network approaches.

## **9.2 Data Collection**

This empirical study was built on the data collected by Lin [1993]. The sample consisted of firms that were either in operation or went bankrupt between 1974 to 1985. 88 failed and 176 nonfailed firms which were listed in the International Stock Exchange Official Year Book were obtained from a Datastream database for the year prior to failure for bankrupt firms and a corresponding year period for each nonfailed firm. To avoid the effect of the extraneous variance, one failed firm was matched with two nonfailed firms based on industry and total assets. Whether this data set is used on full scale or is used to generate multiple subsamples depends on the relevant tests.

## **9.3 Descriptive Statistics of Data Set**

The features of the original data set are examined by descriptive statistics for a total of 264 sample companies, bankrupt firms and nonbankrupt firms respectively. In this study the

predictor variables which we proposed in the preceding chapter were defined as the following 12 financial ratios

R1: NI/TA, net income to total assets ratio

R2: NW/TA, net worth to total assets ratio

R3: NS/TA, net sales to total assets ratio

R4: TL/TA, total liability to total assets ratio

R5: RE/TA, retained earnings to total assets ratio

R6: QA/CL, quick assets to current liabilities ratio

R7: TD/TA, debt to assets ratio

R8: CA/CL, current assets to current liabilities ratio

R9: TC/TS, cash flow to net sales ratio

R10: TS/WC, sales to working capital ratio

R11: QA/TS, quick assets to sales ratio

R12: IC/EB, the ratio of interest payment to earnings before interest and tax

The financial ratios above were selected on the basis of the theories of bankruptcy developed in Chapter Eight, and comprised the set of independent variables successfully used in the construction of prediction models in published research. The mean, standard deviation, skewness, kurtosis of these ratios; the Shapiro-Wilk statistics for univariate normality, and the Mardia test for multivariate normality are provided for all the sample companies and the two different groups.

Additionally, the correlation analysis between ratios is implemented in order to investigate the interrelationship among these explanatory variables. The variance-covariance matrices are computed for each of two groups to examine if the equality of group dispersion across all variables is achieved. Finally, the Mann-Whitney U test is used to assess the statistical significance of the differences in the means of the 12 financial ratios between the bankrupt and nonbankrupt firms.

### 9.3.1 Tests for Normality

#### Univariate Normality

The results of the descriptive statistics and the test of the normal distribution assumption for each of the ratios are presented in Table 9.3.1. It reveals that all ratios are skewed, and that most of them have a positive sign. Only R1, R3 and R10 have a negative skew. A positive sign means that the distribution of the ratios is skewed to the right. Since normal distribution is symmetric and unimodal, it has a skewness index of zero and a kurtosis index of three. As can be seen from Table 9.3.1, however, the kurtosis statistics associated with the 12 financial ratios selected indicate much larger than 3 in 9 of 12 ratios and smaller than 3 in the remaining 3 ratios. Based on this investigation, the selected financial ratios are initially identified not to be univariate normal. In the more strict normal testing procedure, the Shapiro-Wilk statistics is computed. This test has been proven to be better than the alternative Kolmogorov-Smirnov test, especially for a small sample size [Dunn and Clark, 1987]. The results show that apart from R1, R3, and R4 for bankrupt firms, all other cases significantly reject the null hypothesis of normality at  $p < 0.01$  level.

#### Multivariate Normality

In addition to the univariate test for normality, we are also interested in the multivariate test for normality. For a bivariate distribution with random variables  $(x_1, x_2)$ , in large samples the ordered distances  $m_i^2$ ,  $i = 1, 2, \dots, n$ , can be compared to the  $\chi^2$  distribution,  $\chi_{(1-\alpha);2}^2$ , where  $(1-\alpha_i) = (i-0.5)/n$ . The distance  $m^2$  is the squared Mahalanobis distance between  $(x_1, x_2)$  and  $(\bar{x}_1, \bar{x}_2)$  given by

$$m_i^2 = \frac{1}{(1-r^2)} \left[ \frac{(x_{1i} - \bar{x}_1)^2}{s_{x_1}^2} + \frac{(x_{2i} - \bar{x}_2)^2}{s_{x_2}^2} - \frac{2r(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{s_x} \right]$$

A plot of the points  $(m_i^2, \chi_{(1-\alpha);2}^2)$  should yield a straight line.

This plotting technique can be extended to the multivariate normal case by computing the squared Mahalanobis distances for the  $n$  multivariate observations

$$m_i^2 = (\mathbf{X}_i - \bar{\mathbf{X}})' \mathbf{S}^{-1} (\mathbf{X}_i - \bar{\mathbf{X}}) \quad i = 1, 2, \dots, n.$$

The ordered distances  $m^2$  are then plotted against the  $\chi^2$  distribution percentiles,  $\chi^2_{(1-\alpha);p}$ , where  $(1 - \alpha_i) = (i - 0.5)/n$ ,  $i = 1, 2, \dots, n$ ; and  $P$  is the number of variables.

Tests for multivariate normality can also be measured by multivariate skewness and kurtosis. The Mardia [1970] sample measures of multivariate skewness and kurtosis are given by

$$\hat{\gamma}_{1p} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n m_{ij}^3 \quad \text{and} \quad \hat{\gamma}_{2p} = \frac{1}{n} \sum_{i=1}^n m_i^4$$

where

$$m_i^2 = (\mathbf{X}_i - \bar{\mathbf{X}})' \mathbf{S}^{-1} (\mathbf{X}_i - \bar{\mathbf{X}}) \quad \text{and} \\ m_{ij} = (\mathbf{X}_i - \bar{\mathbf{X}})' \mathbf{S}^{-1} (\mathbf{X}_j - \bar{\mathbf{X}})$$

In large samples from a multivariate normal,  $n\hat{\gamma}_{1p}/6$  has a  $\chi^2$  distribution with  $(p+1)(p+2)/6$  degrees of freedom, and  $\hat{\gamma}_{2p}$  is normally distributed with mean  $p(p+2)$  and variance  $8p(p+2)/n$ .

Since the values of  $\chi^2$  corresponding to  $(i-0.5)/n$  are difficult to obtain, the Mardia test was used to examine the multivariate normality. The values of the measures of multivariate skewness and multivariate kurtosis in our sample were determined to be  $\hat{\gamma}_1 = 46.52$  and  $\hat{\gamma}_2 = 110.67$  respectively. The value of  $n\hat{\gamma}_1/6 = 2046.88$ , when compared to a 364 degree of freedom  $\chi^2$ , yielded a p-value less than 0.000. For  $\hat{\gamma}_2$ , the Z value was determined to be 6.78, once again suggesting the extremely small p-value. The two measures therefore show that 12 financial ratios in the empirical data are not multivariate normal.

Table 9.3.1 Descriptive Statistics for Each of 12 Financial Ratios

Descriptive Statistics		All Firms	Bankrupt Firms	Nonbankrupt Firms
R1:NI/TA				
Mean	-0.52992	-9.98216	4.196193	4.196193
Standard deviation	10.69493	13.39249	3.916727	3.916727
Skewness	-3.56242	-3.41526	1.943559	1.943559
Kurtosis	25.24047	19.12853	15.67626	15.67626
Shapiro-Wilk statistics	0.76584	0.763918	0.885397	0.885397
R1:NI/TA				
Mean	-0.52992	-9.98216	4.196193	4.196193
Standard deviation	10.69493	13.39249	3.916727	3.916727
Skewness	-3.56242	-3.41526	1.943559	1.943559
Kurtosis	25.24047	19.12853	15.67626	15.67626
Shapiro-Wilk statistics	0.76584	0.763918	0.885397	0.885397
R2:NW/TA				
Mean	0.475114	0.316136	0.554602	0.554602
Standard deviation	0.222133	0.14842	0.210096	0.210096
Skewness	2.760227	0.529882	4.416680	4.416680
Kurtosis	24.05105	0.199628	39.48893	39.48893
Shapiro-Wilk statistics	0.874674	0.971553	0.767207	0.767207
R3:NS/TA				
Mean	1.58572	1.426705	1.665227	1.665227
Standard deviation	0.819861	0.537018	0.920726	0.920726
Skewness	3.071028	0.164339	3.083622	3.083622
Kurtosis	15.59837	0.288610	13.41087	13.41087
Shapiro-Wilk statistics	0.774537	0.979141	0.736239	0.736239
R4:TL/TA				
Mean	56.13955	72.46705	47.97580	47.97580
Standard deviation	19.55902	15.62883	15.88744	15.88744
Skewness	0.400173	0.301383	0.630347	0.630347
Kurtosis	0.247037	1.921029	1.43137	1.43137
Shapiro-Wilk statistics	0.974686	0.984281	0.973177	0.973177
R6:QA/CL				
Mean	0.835303	0.533295	0.986307	0.986307
Standard deviation	0.511296	0.292711	0.530556	0.530556
Skewness	1.833683	4.056895	1.609146	1.609146
Kurtosis	4.862643	26.03515	4.042802	4.042802
Shapiro-Wilk statistics	0.858331	0.715003	0.885814	0.885814

Prob<W - Associated probabilities, for testing the null hypothesis that the data comes from a normal distribution  
 \*denotes significantly rejects the null hypothesis of normality at  $p < 0.01$  level

**Table 9.3.2 The Correlation Analysis between 12 Financial Ratios**

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0

	R1	R2	R3	R4	R5	R6
R1	—	0.36524 0.0001	0.18242 0.0029	-0.34285 0.0001	0.84964 0.0001	0.33339 0.0001
R2		—	0.18294 0.0028	-0.80176 0.0001	0.38688 0.0001	0.51669 0.0001
R3			—	0.14644 0.0173	0.09274* 0.1329	-0.08186* 0.1848
R4				—	-0.38998 0.0001	-0.58861 0.0001
R5					—	0.30807 0.0001
R6						—

	R1	R2	R3	R4	R5	R6
R7	-0.37648 0.0001	-0.69159 0.0001	0.14643 0.0173	0.92975 0.0001	-0.36046 0.0001	-0.57903 0.0001
R8	0.36014 0.0001	0.52986 0.0001	-0.15363 0.0124	-0.67520 0.0001	0.40008 0.0001	0.72691 0.0001
R9	-0.00188* 0.9758	0.15051 0.0144	-0.22262 0.0003	-0.18232 0.0029	0.03093* 0.6168	0.50708 0.0001
R10	0.14566 0.0179	0.16821 0.0062	0.21998 0.0003	-0.14200 0.0210	0.09637 0.1183	0.02812* 0.6492
R11	-0.11074* 0.0724	0.07500* 0.2246	-0.35788 0.0001	-0.13385 0.0297	-0.06726* 0.2762	0.49213 0.0001
R12	-0.09777* 0.1130	-0.02094* 0.7348	-0.03708* 0.5486	-0.04768* 0.4405	-0.09639* 0.1182	0.02520* 0.6835

\* denotes the situation can not reject the null hypothesis Ho: Rho=0

Table 9.3.3 The Variance-Covariance Matrices for Bankrupt and Nonbankrupt Firms

GROUP = Nonbankrupt Firms							GROUP = Bankrupt Firms						
Variable	R1	R2	R3	R4	R5	R6	Variable	R1	R2	R3	R4	R5	R6
R1	15.3408						R1	179.359					
R2	0.3815	0.0441					R2	-0.424	0.022				
R3	0.4801	0.0326	0.8477				R3	1.589	-0.003	0.288			
R4	-11.8047	-2.3673	4.6627	252.4106			R4	41.100	-1.830	1.659	244.26		
R5	14.3877	0.3317	-0.3625	-22.5325	28.5431		R5	154.747	-0.143	0.873	37.472	228.768	
R6	0.5664	0.0501	-0.0779	-4.8186	0.3994	0.2815	R6	0.041	0.004	-0.020	-0.620	-0.028	0.086
R7	14.3637	-2.0475	5.0846	225.1544	-22.9481	-4.8734	R7	27.418	-1.681	2.371	304.557	66.019	-1.498
R8	0.4496	0.0640	-0.1962	-7.4782	1.0332	0.2675	R8	0.284	0.011	-0.016	-1.813	0.255	0.062
R9	0.0543	0.0079	-0.0455	-0.8312	0.0701	0.0616	R9	-0.472	-0.000	-0.032	-0.015	-0.315	0.032
R10	3.5036	0.3239	0.5419	17.3287	3.0291	-2.2083	R10	-0.642	0.718	2.364	-55.686	-74.199	0.645
R11	0.0238	0.0053	-0.0797	-0.8407	0.0233	0.0739	R11	-0.947	0.002	-0.063	-0.299	-0.695	0.043
R12	0.0032	0.0002	-0.0003	-0.0345	0.0059	0.0011	R12	-0.021	-0.000	-0.000	-0.019	-0.027	-0.001

Variable	R7	R8	R9	R10	R11	R12	Variable	R7	R8	R9	R10	R11	R12
R1							R1						
R2							R2						
R3							R3						
R4							R4						
R5							R5						
R6							R6						
R7	233.836						R7	489.924					
R8	-7.8052	0.5602					R8	-2.998	0.145				
R9	0.8781	0.0492	0.0559				R9	-0.135	0.018	0.023			
R10	35.1031	-4.1653	-0.4200	692.5219			R10	-65.472	3.818	0.209	2622.251		
R11	-0.9094	0.0502	0.0605	-0.8918	0.0722		R11	-0.509	0.013	0.028	-0.250	0.046	
R12	-0.0390	0.0011	0.0001	-0.0234	0.0001	0.0001	R12	-0.018	-0.000	-0.000	0.025	-0.000	0.000



### 9.3.2 Correlation Analyses and Variance-Covariance Matrices

In order to examine the characteristics of sample data further, the correlation analysis is computed and presented in Table 9.3.2. As it indicates, at  $p < 0.01$  level, most financial ratios are significantly correlated to each other apart from R5, R11 and R12. Additionally, the variance-covariance matrices for both bankrupt and nonbankrupt firms have been investigated. Table 9.3.3 shows the matrices for these two groups. It is obvious from Table 9.3.3 that failing and nonfailing firms do not have equal covariance structure. For instance, in nonbankrupt firms, the variances for R1 and R5 are 15.3408 and 28.5431, respectively, but in bankrupt firms, the corresponding variances are 179.359 and 228.768, respectively. Pairwise comparisons of the variance-covariance matrices clearly indicate that bankrupt and nonbankrupt groups have different dispersion for the 12 financial ratios selected. Consequently, as shown in previous studies, both multivariate normality and equal variance-covariance matrices assumptions are hardly obtained in the real word in bankruptcy prediction models.

**Table 9.3.4 The Mann-Whitney U test for Differences in the Means of 12 Financial Ratios between Failing and Nonfailing Firms**

	Mean Score in Failing Firms	Mean Score in Nonfailing Firms	Z-value	P >  Z
R1	54.38	171.56	-11.75	0.0001*
R2	69.41	164.04	-9.94	0.0001*
R3	121.85	137.82	-1.60	0.1093
R4	197.85	99.83	9.83	0.0001*
R5	56.67	170.41	-11.41	0.0001*
R6	73.53	161.99	-8.87	0.0001*
R7	202.22	97.64	10.49	0.0001*
R8	70.30	163.6	-9.36	0.0001*
R9	99.56	148.97	-5.22	0.0001*
R10	118.76	139.37	-2.07	0.0388*
R11	127.43	135.03	-0.76	0.4459
R12	154.15	121.68	3.77	0.0001*

\*denotes that there is significant difference in the mean value between bankrupt and nonbankrupt firms at  $p < 0.05$  level

### **9.3.3 The Difference in the Means of Twelve Financial Ratios between Failing and Nonfailing Firms**

Since the explanatory variables are not multivariate normal and the group dispersions are not homogeneous, the nonparametric Mann-Whitney U test is utilised to examine the mean location differences between bankrupt and nonbankrupt groups for all 12 selected financial ratios. The results of Mann-Whitney U test are presented in Table 9.3.4.

As can be seen from Table 9.3.4, all financial ratios selected except for the R3 and R11 exhibit significant differences between the two groups' mean levels at a  $p < 0.05$  level in terms of univariate assessment. However, these two ratios may have good multivariate discrimination when combined with other ratios. Further, bankrupt firms have higher mean levels of R4, R7 and R12, compared to nonbankrupt firms.

Generally, the statistically significant results imply that the 12 selected financial ratios possess high discriminating power in distinguish between failing and nonfailing companies. These 12 financial ratios are thus used as the predictor variables for the subsequent experiments for all four classification methods.

## **9.4 The Questions to be Tested and the Corresponding Research Designs**

### **9.4.1 To Verify the Simulation Results in the Real Financial Data Set**

There are five subjects to be explored in the empirical study, the first one is

1. To compare predictive abilities on real data for alternative discriminating approaches and to verify the simulation results.

The data with all 264 companies consisting of 88 failing and 176 nonfailing companies will be analysed by four discriminating techniques. In order to have a contrast with the simulation results, the prior probabilities and misclassification costs of two groups are set to be equal. However, the comparison of their superiority is based on the dominance of the OC (operating characteristic) curve developed by Steele [1995]. This approach avoids the possible inappropriate comparison rule which is determined (according to a particular classification accuracy) by the specific optimal cutoff point (i.e., the specific error cost and prior probability). By using the dominance approach, the ranking is invariant to error costs

and subjective prior probabilities, thus provides a fair comparison among different models. The detail of this process will be presented in Chapter Ten and will be applied to this experiment.

The tool and the computer running procedure used in the simulation study will be employed in this empirical experiment as well. However, since the empirical data has 12 inputs—many more than the bivariate variables in the simulation data—the number of hidden nodes should be carefully determined to achieve not only accurate classification but also generalisation.

Following up the method of deciding network architecture employed in the simulation study, the number of hidden nodes is also determined by the Cascade-correlation algorithm (Cascor). The whole sample is randomly divided into two subsamples of equal size. One is the training data set, the other is the testing data set. In order to have reliable results, a more accurate approach suggested in Berry and Trigueiros [1993] is used here. The training set is randomly subdivided into a number of two parts, A and B. Training set A is for learning optimal network structure and training set B is for testing whether the structure of the trained network is an optimal structure. Training set A is used to do weight updating through Cascor. The results obtained from training sets A are examined in training sets B. The topology that gives best performance is then selected and the true generalisation ability of the network topology can then be checked on the as yet unused testing set. The optimal topology found in the empirical data is 12-9-1. This architecture will also be used in the other four experiments of the empirical study.

#### **9.4.2 To Evaluate the Influence of Different Sample Size**

The second subject to be explored in this empirical study is

2. To test the effects of the different sample sizes on predictive ability for the four methods. Many researchers have been concerned about the number of observations needed to develop a reliable model. For example, Stam and Jones [1990] studied small and medium sample sizes in discriminant analysis, and concluded that better classification performance was obtained as the size of sample increased. We also want to know if increasing the sample size to a large sample size will further increase classification performance in other approaches. There is no general rule to determine what size will be appropriate for each

different case. Most studies choosing their sample size based on the availability of data. Although we cannot establish a methodology for choosing sample size, an understanding of the effects of sample size on prediction performance is at least necessary if the model developed is to be reliable. Thus, small, medium and large training and validation sample sizes of 30, 60, 120 observations respectively will be employed to build the classification accuracy insight for each of the four models.

The research design will generate multiple subsamples from the original data set consisting of 88 failed and 176 nonfailed firms by varying sample size with equal observations in each group. For each case, 20 different learning and validation tests are obtained individually to study the impact of this factor. To avoid confounding the interpretation of test results, the misclassification costs and the prior probabilities are set equal.

#### **9.4.3 To Investigate the Influence of Choice-Based Bias and WCOP Procedure**

The third subject to be explored in this empirical study is

3. To test the effects of different levels of sample frequency rates on predictive abilities, and to assess if the choice-based sampling bias could be eliminated by comparing the differences between incorporating the WCOP adjustment procedure and not incorporating adjustment procedure among these four techniques.

The sample frequency rate is defined as the proportion of bankrupt firms to nonbankrupt firms in the sample (i.e., base rate in the sample). This is referred to matching criteria in prior studies. The match-pairs approach is a way to control the extraneous factor which may confuse the research results. Beaver [1966] recommended that the matching design should be selected to provide a control over factors which might obscure the connection between financial ratios and business failure. Lev [1974] also noted that the paired-sample method permits researchers "to control for various factors that are believed to be unrelated to the phenomenon investigated" (p.141). One empirical study [Altman, 1968] proved that matched and unmatched sampling design with same financial ratios led to significant differences in prediction accuracy. For example, the size and feature of a firm are generally regarded as having some influences on the probability of business failure. Therefore, most

studies attempted to eliminate these effects by using a matched sample based on the book value of total assets, sales or industry, etc.

The question is: if a researcher uses one to one matched sampling (equal base rate), while other researchers choose different matched criteria, say one to five or one to ten from the same population, is there a significant difference in classification accuracy between them when other factors remain fixed? If there are some impacts on classification accuracy, what kind of impact is caused, the Type I error rate, Type II error rate or overall error rate? and how deep is it? or how sensitive is it for a particular technique?

On the other hand, when the base rate in a sample is not equal to the proportion of two groups in the population, it probably causes a choice-based sample bias of both the parameter and probability estimates. According to the previous studies by Zmijewski [1984], Dopuch, Holthausen and Leftwich, [1987], and Manski and Lerman [1977], if an appropriate adjustment scheme is used, this bias could be eliminated. Based on the above idea, the adjusted procedure WCOP (weighted cutoff point) was proposed and will be applied to the four discriminating techniques in order to evaluate if such bias can be removed or mitigated.

This study will be divided into two small experiments. First, the samples are conducted on no-adjustment process using data sets with different sample frequency rates. The second has the same procedure but incorporates the WCOP adjustment in order to provide contrasts with the former ones. Six choice-based samples are drawn from the full data set. The six samples each contains 15 bankrupt firms but different (increasing) numbers of nonbankrupt firms, selected randomly from the 176 available nonbankrupt firms in the full data set. The number of nonbankrupt firms in the six estimation samples is 15, 30, 60, 75, 120 and 150. The resulting proportions of failing to nonfailing firms are thus 1:1, 1:2, 1:4, 1:5, 1:8 and 1:10. The choice-based sampling issue will be assessed empirically by comparing the weighted and unweighted results across these six cases.

The assumption of equal misclassification costs is made. However, the prior probability assumes 0.09 (i.e., 1:10 proportion of bankruptcy in the population). This proportion is selected in order to contrast with the smallest sample selection probability (1:10), and is expected to reveal the elimination of choice-based sampling bias when the WCOP procedure is used. The research design is depicted as follows

**Table 9.4.1 The Research Design of Weighted and Unweighted Procedure on Six Choice-Based Estimation Samples**

Choice-Based Estimation Sample						
	15:15	15:30	15:60	15:75	15:120	15:150
Unweighted						
Bankrupt	Type I	Error	Rate	Across	the Six	Samples
Nonbankrupt	Type II	Error	Rate	Across	the Six	Samples
Overall	Overall	Error	Rate	Across	the Six	Samples
Weighted						
Bankrupt	Type I	Error	Rate	Across	the Six	Samples
Nonbankrupt	Type II	Error	Rate	Across	the Six	Samples
Overall	Overall	Error	Rate	Across	the Six	Samples

For the WCOP procedure, as we proposed in Chapter Eight, if prior probabilities in the population, the base rates in the sample and the two kinds of misclassification costs are all considered, the optimum cut-off point  $I^*$  depends on the estimates of three parameters: (1) the judgements on the ratio of two kinds of misclassification cost; (2) the ratio of prior probability of nonfailing and failing firms in the population; (3) the ratio of the proportion of failing and nonfailing firms used in sample.

The optimal cut-off point  $I^*$  which minimises the total adjusted cost (AC) can be obtained when the following equation is solved.

$$\frac{f(I^*/B)}{f(I^*/N)} = \frac{C_{II}}{C_I} \times \frac{1 - \alpha_p}{\alpha_p} \times \frac{\alpha_s}{1 - \alpha_s} = K \quad (9.4.1)$$

where

$\alpha_s$  = the proportion of bankruptcy in the sample

$\alpha_p$  = the proportion of bankruptcy in the population

$C_I$  = the cost of misclassifying a bankrupt firm as a nonbankrupt firm, i.e. misclassification cost of Type I error

$C_{II}$  = the cost of misclassifying a nonbankrupt firm as a bankrupt firm, i.e. misclassification cost of Type II error

In the MDA model, or the Probit model, we assume the distribution of I is normal distribution, then I\* can be found by solving the following equation (see Chapter 8)

$$\begin{aligned} & (I^* - \mu_2)^2 / 2\sigma_2^2 - (I^* - \mu_1)^2 / 2\sigma_1^2 \\ & = \ln(\sigma_1 / \sigma_2) + \ln[\alpha_s / (1 - \alpha_s)] + \ln[(1 - \alpha_p) / \alpha_p] + \ln(C_{II} / C_I) \quad (9.4.2) \end{aligned}$$

where

$\mu_1$  = the means of Z scores for bankrupt firms

$\mu_2$  = the means of Z scores for nonbankrupt firms

$\sigma_1$  = the standard deviation of Z scores for bankrupt firms

$\sigma_2$  = the standard deviation of Z scores for nonbankrupt firms

The estimates of  $\mu_1$ ,  $\mu_2$ ,  $\sigma_1$  and  $\sigma_2$  are sample means and variances of the Z scores for two groups, that is,  $\bar{x}_1$ ,  $\bar{x}_2$ ,  $s_1$  and  $s_2$  for the bankrupt group and nonbankrupt group, and can be empirically estimated from a sample of bankrupt firms and a sample of nonbankrupt firms. Applying the 12 predictors of 264 companies to the MDA discriminant function, we obtain the Z scores on each of all 264 observations and their descriptive statistics by respective groups. They were shown in Appendix V.

The estimates of parameters  $\mu_1$ ,  $\mu_2$ ,  $\sigma_1$  and  $\sigma_2$  are

$$\begin{aligned} \bar{x}_1 &= 3.6237575 & s_1 &= 3.5765666 \\ \bar{x}_2 &= -3.6243104 & s_2 &= 2.1193865 \end{aligned}$$

Inserting the above estimates into equation (9.4.2), the optimal cutoff point I\* can be achieved by solving the numerical root of the equation when the  $\alpha_s$ ,  $\alpha_p$  values are given. The  $\alpha_s$  values designed in this experiment are 1/2, 1/3, 1/5, 1/6, 1/9 and 1/11 respectively; under the assumption of prior probability of bankruptcy  $\alpha_p = 0.09$  (i.e., 1/11), their corresponding optimal cutoff points I\* are reported in Table 9.4.2.

Table 9.4.2 The Optimal Cutoff Points in MDA Approach

Sample Base Rate	$\alpha$ , value	Without Adjustment	With WCOP
15:15	1/2	-4.010781	1.560410
15:30	1/3	-4.010781	1.017340
15:60	1/5	-4.010781	0.437414
15:75	1/6	-4.010781	0.241536
15:120	1/9	-4.010781	-0.188075
15: 150	1/11	-4.010781	-4.010781

As a matter of fact, in SAS discriminant analysis, a statement PRIORS in DISCRIM procedure is provided to cope with the unequal prior probability situation. Accordingly, the WCOP procedure in MDA can also be constructed using different k value as same weight in PRIOR statement.

In terms of Logit approach, Index I used here is referred to as the intermediate value Z utilised in calculation of p (conditional probability). That is,  $Z = \beta'X$ ,  $p = 1/(1+\exp(-Z))$ , where  $\beta'$  is the vector of estimated coefficients of predictor variables, X is the vector of predictors. For logistic regression, it implies that these Z values have a logistic distribution. As it was discussed before, if the specific probability density function (p.d.f.) is applied to equation (8.3.10) given a fixed k value, then the optimal cutoff point  $I^*$  is achieved by solving the root of equation using numerical iteration methods such as Newton's iteration. The detailed steps for Logit procedure are then shown below

$$f(I) = \frac{\frac{1}{b} \exp(-(\frac{I-a}{b}))}{\left[1 + \exp(-\frac{I-a}{b})\right]^2}$$

I is a variable from logistic distribution  
a any value,  $b > 0$

$$F(I) = \frac{1}{1+e^{-I}}$$

F is the cumulative distribution function of I

$$E(I) = a$$

Mean of logistic function

$$\text{Var}(I) = \frac{b^2 \pi^2}{3} \rightarrow b = \frac{\sigma}{\pi} \sqrt{3}$$

Variance of logistic function



Applying Logit procedure, the intermediate values  $Z = \beta'X$ , the conditional probabilities  $p$  of each company and their descriptive statistics for failing and nonfailing groups are displayed in Appendix VI and VII.

The estimates of parameters  $\mu_1$ ,  $\mu_2$ ,  $\sigma_1$  and  $\sigma_2$  in terms of intermediate  $Z$  values are calculated as

$$\begin{aligned}\bar{x}_1 &= 7.6467750 & s_1 &= 8.2426481 \\ \bar{x}_2 &= -6.2517737 & s_2 &= 3.9908628\end{aligned}$$

Thus,

$$\begin{aligned}a_1 &= 7.6467750 & b_1 &= 4.54441 \\ a_2 &= -6.2517737 & b_2 &= 2.20028\end{aligned}$$

Here, we set equal misclassification costs of Type I and Type II error, and the prior probability of bankruptcy  $\alpha_p$  is assumed to be 0.09. That is, under WCOP procedure, the equation (9.4.1) becomes

$$\frac{f(I^*/B)}{f(I^*/N)} = \frac{1-\alpha_p}{\alpha_p} \times \frac{\alpha_s}{1-\alpha_s} = \frac{1-\frac{1}{11}}{\frac{1}{11}} \times \frac{\alpha_s}{1-\alpha_s} = 10 \frac{\alpha_s}{1-\alpha_s} = K \quad (9.4.3)$$

If we don't take WCOP adjustment procedure into account, the equation (9.4.1) will be

$$\frac{f(I^*/B)}{f(I^*/N)} = 1 = K \quad (9.4.4)$$

Then, replace the distribution of  $f(I/B)$  and  $f(I/N)$  into equation (9.4.3) and (9.4.4) with the logistic p.d.f, we get

$$\frac{\frac{1}{b_1} \exp(-(\frac{I^*-a_1}{b_1}))}{(1+\exp(-(\frac{I^*-a_1}{b_1}))^2} \div \frac{\frac{1}{b_2} \exp(-(\frac{I^*-a_2}{b_2}))}{(1+\exp(-(\frac{I^*-a_2}{b_2}))^2} = K \quad (9.4.5)$$

Insert  $a_1$ ,  $a_2$ ,  $b_1$ ,  $b_2$ , and  $k$  value into equation (9.4.5), the optimal cutoff  $I^*$  then is obtained by finding the numerical root using the Newton's iteration. For example, for one to one matched criterion  $\alpha_s$  is 0.5, thus  $k = 10$ , given the above  $a_1$ ,  $a_2$ ,  $b_1$ ,  $b_2$ , the  $I^*$  approximation is 2.0547152. However, this optimal cutoff  $I^*$  is in the form of intermediate value, we transform it through logistic distribution function (c.d.f.) in order to compare with the conditional probabilities obtained from Logit model. The various  $k$  values and their corresponding optimal cutoff points based on the intermediate value  $Z$  and transformed value  $p$  (probability) were calculated and presented in Table 9.4.3 and Table 9.4.4.

Table 9.4.3 The Various K Values in Logit Approach

Sample Base Rate	K value Without Adjustment	K value With WCOP
15:15	1	10
15:30	1	10/2
15:60	1	10/4
15:75	1	10/5
15:120	1	10/8
15:150	1	1

Table 9.4.4 The Optimal Cutoff Points in Logit Approach

	Optimal cutoff intermediate	$I^*$ based on the value $Z$	Optimal cutoff transformed	$I^*$ based on the value $p$
Sample Base Rate	Without Adjustment	With WCOP	Without Adjustment	With WCOP
15:15	-0.5206249	2.0547152	0.372706	0.886423
15:30	-0.5206249	1.2570709	0.372706	0.778521
15:60	-0.5206249	0.4843861	0.372706	0.618783
15:75	-0.5206249	0.2387346	0.372706	0.559402
15:120	-0.5206249	-0.2763700	0.372706	0.431344
15:150	-0.5206249	-0.5206249	0.372706	0.372706

We can see from the Table 9.4.4 that when the sample selection probability equals the prior probability in the population, the optimal cutoff point in the unadjustment and WCOP adjustment is identical. In this situation, the choice-based sample bias is eliminated.

Likewise, for a neural network with sigmoid function in the output layer, the predicted output values are between 0 and 1. Since the sigmoid function has the same form as the cumulative logistic function, we treat these predicted values as the transformed values through the logistic distribution function just like the procedure in the Logit model. Hence, we should transform them back to achieve the assumed intermediate  $Z$  values. The back-transformed  $Z$  values and predicted values of our 264 companies and their descriptive statistics by group for GDR and Projection methods are shown in Appendix VIII, IX and Appendix X, XI respectively.

The estimates of  $\bar{x}_1$ ,  $\bar{x}_2$ ,  $s_1$  and  $s_2$  in terms of intermediate  $Z$  values as follows

(1) For the GDR method,

$$\begin{array}{ll} \bar{x}_1 = 3.4168651 & \sigma_1 = 1.7791123 \\ \bar{x}_2 = -4.3621649 & \sigma_2 = 1.5936182 \\ \text{Thus,} & \\ a_1 = 3.4168651 & b_1 = 0.980876 \\ a_2 = -4.3621649 & b_2 = 0.878608 \end{array}$$

(2) For the Proj approach,

$$\begin{array}{ll} \bar{x}_1 = 2.6354295 & \sigma_1 = 1.8319012 \\ \bar{x}_2 = -4.7037655 & \sigma_2 = 1.6768469 \\ \text{Thus,} & \\ a_1 = 2.6354295 & b_1 = 1.009980 \\ a_2 = -4.7037655 & b_2 = 0.924494 \end{array}$$

We apply same computation processes demonstrated before in order to obtain the optimal cutoff points for the GDR and Projection techniques. The results are indicated below

**Table 9.4.5 The Optimal Cutoff Points in GDR Neural Network**

	Optimal cutoff intermediate	I* based on the value Z	Optimal cutoff transformed	I* based on the value p
Sample Base Rate	Without Adjustment	With WCOP	Without Adjustment	With WCOP
15:15	-0.6339589	0.4728378	0.346613	0.616055
15:30	-0.6339589	0.1370237	0.346613	0.534202
15:60	-0.6339589	-0.1958990	0.346613	0.451181
15:75	-0.6339589	-0.3027019	0.346613	0.424897
15:120	-0.6339589	-0.5273559	0.346613	0.371134
15: 150	-0.6339589	-0.6339589	0.346613	0.346613

**Table 9.4.6 The Optimal Cutoff Points in Proj Neural Network**

	Optimal cutoff intermediate	I* based on the value Z	Optimal cutoff transformed	I* based on the value p
Sample Base Rate	Without Adjustment	With WCOP	Without Adjustment	With WCOP
15:15	-1.1516808	0.0219457	0.240182	0.505486
15:30	-1.1516808	-0.3356910	0.240182	0.416857
15:60	-1.1516808	-0.6885688	0.240182	0.334352
15:75	-1.1516808	-0.8015555	0.240182	0.309693
15:120	-1.1516808	-1.0390327	0.240182	0.261337
15: 150	-1.1516808	-1.1516808	0.240182	0.240182

These optimal cutoff points can then be used to determine the classification rule and to achieve the corresponding classification accuracy. Therefore, the comparative analysis between without WCOP adjustment and with WCOP adjustment can proceed.

The existence of a choice-based sampling bias will be examined by comparing the unweighted and weighted results across the six estimation samples. If a bias exists, then there should be some functional relationship between the various sample base rates and the

individual group classification results. On average, lower bankrupt firm sample frequency rates result in lower bankrupt firm correct classification, and the lower bankrupt firm sample frequency rates can lead to higher nonbankrupt firm classification accuracy. Pearson correlation coefficients are used in order to identify if this bias is present and if it can be eliminated by WCOP.

#### **9.4.4 To Conduct a Sensitivity Analysis of Optimal Cutoff Points to Different Ratios of Misclassification Costs**

The fourth subject addressed in the research is

4. To make a sensitivity analysis of optimal cutoff points to misclassification costs of Type I and Type II errors.

Type I error is the misclassification of a failing firm as a nonfailing firm, and Type II error is the misclassification of a nonfailing firm as a failing firm. When a failing firm is predicted as a nonfailing firm, investors may lose their whole investment. On the other hand, when a nonfailing firm is predicted as a failing firm, investors lose the dividend income and capital gain that would otherwise be obtained. Generally, the misclassification cost of a Type I error is much higher than that of the Type II error. Therefore, when the optimal cut-off point is determined without considering the loss functions of Type I and Type II errors, the results may not be optimal for a user to apply the bankruptcy prediction model. Koh [1992] undertook a study which is one of limited research to explore this issue, yet he only made a sensitivity analysis in the Logit model. Hence, the objective of this experiment is to extend Koh's study to investigate the sensitivity of optimal cutoff points to the misclassification costs of Type I and Type II errors for all four discriminating techniques.

In this study the Type I error cost ( $C_I$ ) and Type II error cost ( $C_{II}$ ) were not directly measured. Instead, the expected misclassification costs of using the model were computed under alternative assumptions about the relative misclassification costs of Type I and Type II errors (i.e. the ratio  $C_I$  to  $C_{II}$  or  $C_I:C_{II}$ ).

Following Koh's study [1992], the  $C_I$  to  $C_{II}$  was ranging from 1:1 to 500:1. The ratio 1:1 is a lower limit since the misclassification cost of a Type I error is expected to be higher than

that of a Type II error. Further, ratio 500:1 is believed to sufficiently represent the situation in which Type I error cost is much larger than Type II error cost.

The prior probabilities of two groups are set to be equal in order to reveal only the influences of changing ratios of misclassification costs themselves. As a matter of fact, for different prior probabilities, the analysis can be made by just changing  $k$  value.

In this experiment the optimal cutoff points were also obtained by solving the equation (9.4.1) for four discriminating techniques. The results among the four methods were plotted in order to demonstrate the respective and relative degree of change in optimal cutoff points to corresponding changes in misclassification cost ratios.

#### **9.4.5 To Assess the Influence of Different Base Rates between the Training and Testing Samples**

The last subject we are interested in is

5. To test the effects on generalisation ability in terms of differences in base rates between the training and testing samples.

The decision maker in the real world may not have control over the composition of historical data necessary for predictive model development, and it is meaningful to know whether a classification model built using a training sample with a certain base rate still works when the prior probabilities in the test population becomes very different.

In order to study the effects of this proportion on the predictive performance of the four techniques, we created three proportions for each of the training and testing set compositions. The levels are 1:1, 1:5 and 1:9. A full two-factor design is used, and nine different experimental cells are thus produced. Within each cell 20 different training-testing set pairs are generated. These pairs contain unique firms, i.e. no overlap is allowed. We fix the sample size at a reasonable level, 60 bankrupt and 60 nonbankrupt firms. The prior probabilities of two groups and two types of misclassification cost are both assumed to be equal. The research design is indicated in the following table

**Table 9.4.7 The Research Design of Investigating the Influence of Different Base Rate between Training and Validation Samples on Predictive Ability**

Sample size 120	Testing set base rate	Testing set base rate	Testing set base rate
Training set base rate 1:1	60/60 60/60 replicate 20	60/60 20/100 replicate 20	60/60 12/108 replicate 20
Training set base rate 1:5	20/100 60/60 replicate 20	20/100 20/100 replicate 20	20/100 20/100 replicate 20
Training set base rate 1:9	12/108 60/60 replicate 20	12/108 20/100 replicate 20	12/108 12/108 replicate 20

## 9.5 Summary and Conclusions

The research designs and experiment methodologies of each of the empirical subjects have been presented in this chapter. They include

- (1) To make comparisons in the classification accuracy of the bankruptcy prediction model on real data for alternative discriminating approaches
- (2) To explore the effects of the different sample sizes on predictive ability for the four discriminating methods.
- (3) To test the effects of different base rate levels on the predictive abilities, and to assess if the choice-based sampling bias could be eliminated through comparing the differences between incorporating the adjustment WCOP and not incorporating adjustment procedure among these four techniques.
- (4) To make the sensitivity analysis of optimal cutoff points to misclassification costs of Type I and Type II errors.
- (5) To investigate the effects on generalisation in terms of differences in base rate between training and testing samples.

This empirical study attempts to fully discuss various facets of bankruptcy prediction models. Since the complicated interrelationship and characteristics between independent variables can not be described exhaustively in a simulation process, it may be more practical to use real data to make a comparison and inference about the predictive ability. The results can also be used as a way of verifying those obtained from the simulation study. In addition, the exploration related to the impacts of some factors we frequently experienced in reality, such as the sample size selection, the choice-based sampling bias derived from the matched criterion design, the sensitivity of optimal cutoff points to unequal misclassification costs, and the generalisation issue are also included. The detailed results of the five experiments and their analyses will be discussed in the subsequent chapter.



## Chapter Ten

### RESULTS OF THE EMPIRICAL STUDY

#### 10.1 Introduction

This chapter reports the results of five experiments and analyses the hypotheses tested using real financial data. The five experiments include comparing the classification accuracy for MDA, Logit, GDR and Proj methods; assessing the impact of sample size on predictive ability; evaluating the influence of choice-based sampling bias; testing the sensitivity of optimal cutoff points to two kinds of misclassification costs; and investigating the generalisation ability when using *different proportions* of two groups between the training and testing samples. In addition to the introduction, this chapter is divided into six parts. The first five parts deal with each associated problem described above. The final section offers a conclusion and summary of the discussions.

#### 10.2 The Results between ANNs and STMs Using Real Financial Data

The first comparison in the empirical study involves assessing whether the neural networks could achieve equal or superior predictive power to the conventional statistical discriminating techniques in bankruptcy prediction based on real financial data. Further, we are interested to know if the relative importance of independent variables provided by these four methods are significantly inconsistent.

Four relevant hypotheses are proposed for the experiment carried out in this section.

$H_{16}$  : There is no difference in classification performance for the four alternative techniques for the training sample based on real financial data.

$H_{17}$  : There is no difference in classification performance for the four alternative techniques for the testing sample based on real financial data.

$H_{18}$  : There is no significant inconsistency between the simulation results and the empirical results in classification performance for the four alternative techniques.

$H_{19}$  : There is no significant difference in the relative contribution of predictor variables for four alternative techniques

### 10.2.1 Comparison on Classification Accuracy

The real financial data used in the empirical study is from the U.K. Datastream data base, which was described in the previous chapter. In this experiment we divided the whole data set into two halves. One half is for the training data set, and the other half is for the testing data set. Each data set consists of 44 bankrupt and 88 nonbankrupt firms matched on assets and industry. In the neural network, using the 12-9-1 architecture, around 25000 epoch iterations were carried out during the network training phase. The network root mean square (RMS) error and the changes of weights were monitored throughout the training period. We chose this number of iterations as a stop rule because the RMS is small and the change of weights becomes negligible at this level. Prediction results for the training and the testing samples among these four alternative techniques are given in Table 10.2.1

**Table 10.2.1 Comparison in Classification Performance of Four Alternative Techniques for Training and Testing Sample**

Method	Training Data Set			Testing Data Set		
	Type I Error	Type II Error	Overall Error	Type I Error	Type II Error	Overall Error
MDA	9.99%	0	3.03%	13.64%	5.68%	8.33%
Logit	6.82%	2.3%	3.79%	20.45%	7.95%	12.12%
GDR	2.3%	0	0.75%	6.80%	12.50%	10.6%
Proj	0	0	0	9.09%	10.20%	9.85%

The above results were obtained under the assumption of equal prior probability and misclassification costs. One approach that would allow comparability across models involves estimating the relative costs of Type I and Type II errors. In general, the value of Type I error cost and Type II error cost is individual, subjective and context specific. Therefore, previous researchers typically assumed several alternative relative cost ratios, and identified the corresponding classification accuracy or cutoff point under each

assumption. The idea is to determine if either model dominates the other in terms of minimising misclassification costs across a frontier of error rates or optimal cutoff points associated with assumed cost ratios.

Steele [1995] dealt with this issue more subtly and more deeply. He used a dominance function to demonstrate the correct comparison between different models. The concept he developed is described as follows

If the impact of base rate is not considered, then minimising the total cost can be expressed as

$$\text{Min } E(C) = C_{II}(1-\alpha_p) \int_I^{\infty} f(I/N) dI + C_I \alpha_p \int_{-\infty}^I f(I/B) dI \quad (10.2.1)$$

(The definitions of all notations are the same as the equation (8.3.3))

The decision problem indicated by equation (10.2.1) is an unconstrained optimisation problem to choose an optimal cutoff value of  $I^*$ . Steele [1995] solved this problem in a different way by transforming the decision into an equivalent constrained optimisation problem

$$\text{Min } E(C) = aX + bY$$

$$\text{Such that } X = \int_I^{\infty} f(I/N) dI = 1 - F(I/N)$$

$$Y = \int_{-\infty}^I f(I/B) dI = F(I/B)$$

where

$$a = C_{II}(1-\alpha_p) \quad b = C_I(\alpha_p)$$

The constraints are drawn by plotting the curve described by the locus of points  $(1-F(I/N), F(I/B))$  as parameter  $I$  varies across its range. This curve is called the Operating Characteristic (OC) curve by Steele [1995].

Any point on the curve corresponds with the failure probability that a healthy firm has a value from the discriminating model greater than  $I$ , and the nonfailure probability that a bankrupt firm has an index smaller than  $I$ .

The objective function is determined by the prior probability of bankruptcy  $\alpha_p$  and the two misclassification costs  $C_I$  and  $C_{II}$ .

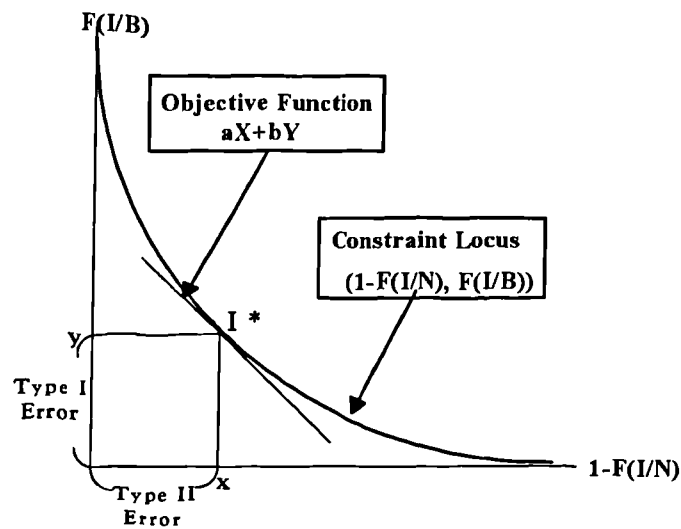


Figure 10.2.1 Discriminant Function  
As an Equivalent Constraint Optimisation

The optimal cutoff point  $I^*$  is found when the gradient of the objective function ( $aX+bY$ ) is equal to the gradient of the constraint. That is,  $-a/b = -f(I^*/B)/f(I^*/N)$ . The co-ordinates at the point of tangency ( $x, y$ ) give the Type II error rate and Type I error rates illustrated in Figure 10.2.1. It can be seen the Type I and Type II error rates are dependent on the objective function and the locus of  $(1-F(I/N), F(I/B))$ . The smaller the prior probability of bankruptcy or Type I error cost relative to Type II error cost, the higher the negative gradient to the objective.

This results in an optimal cutoff point with a high rate of Type I error and a low rate of Type II error. On the other hand, when the whole locus of a model's OC curve (model 1) is nearer to the axes than those of the other's model (model 2), the more predictive power this model has, since whatever the values of  $a$  and  $b$ , (i.e., misclassification costs and prior probabilities), smaller Type I and Type II error rates can both be achieved by this model (model 1). This point is demonstrated in Figure 10.2.2.

In Figure 10.2.2, model 1 always provides lower Type I and Type II errors than model 2, regardless of the change of objective function. In this situation, it is concluded that model 1 is superior to model 2.

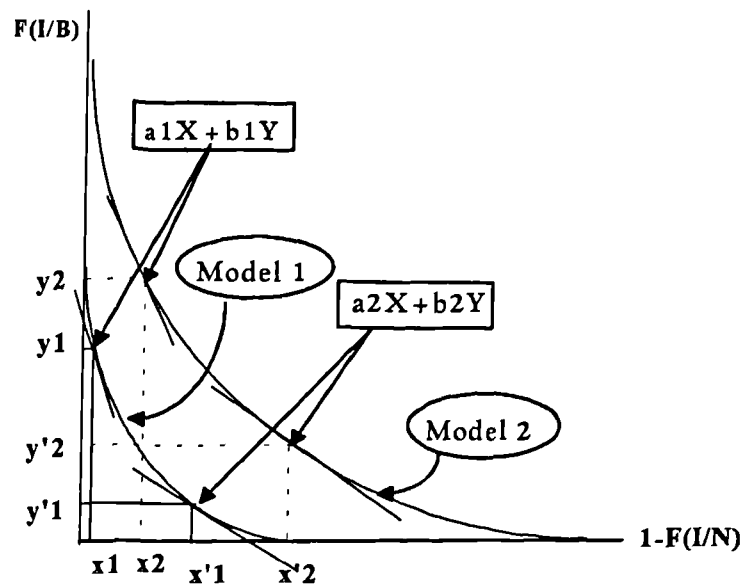


Figure 10.2.2 Situation I of Dominant Discriminating Function

Conversely, if the whole frontier of a model's OC curve cannot dominate the other model's, we cannot conclude which model is superior, since the objective function produces different degrees of probability of Type I and Type II errors for each model. For example, in Figure 10.2.3. when the objective function is  $aX+bY$ , model 1 produces a lower rate of Type I errors, but a higher rate of Type II errors. On the contrary, model 2 offers the higher Type I errors, and lower Type II error rates. In this situation, model 1 and model 2 give ambiguous signals about the relative rankings.

As Steele [1995] pointed out, the previous literature has placed undue emphasis on error rates generated by a specific cutoff point (i.e., a specific ratio of misclassification costs or prior probabilities). Using particular error rates to rank different models is inappropriate owing to the fact that they vary with subjective choices in the objective function.

Following the idea developed by Steele [1995], we decided to assess the predictive capabilities of these four techniques by using the considerations of dominance described above. In order to implement this approach, the distribution of  $f(I/B)$  and  $f(I/N)$  should first be obtained for each of the four methods. Thus the cumulative distribution  $F(I/B)$  and  $F(I/N)$  can be derived. The empirical cumulative distribution for nonbankrupt firms  $F(I_i|B)$  is computed by sorting the index  $I$  in ascending order, and setting  $F(I_i|B) = i/n$ ,

where  $i$  is the rank number of the  $i$ th index  $I$ , and  $n$  is the sample size. For nonbankrupt firms  $1-F(I_i|N)$  is similarly calculated by sorting in descending order. Accordingly, each method has its corresponding OC curve. A fair comparison can then be made.

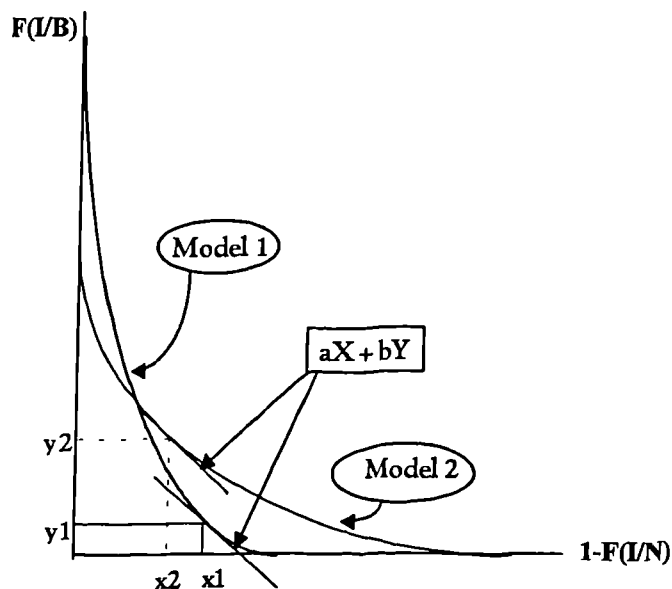


Figure 10.2.3 Situation II of Dominant Discriminating Function

For the MDA method, the frequency distribution and cumulative distribution of  $Z$  scores on bankrupt firms (i.e.,  $f(I/B)$  and  $F(I/B)$ ) as well as on nonbankrupt firms (i.e.,  $f(I/N)$  and  $F(I/N)$ ) are calculated and reported in Table 10.2.2 for the training data set and in Table 10.2.3 for the testing data set.

For the Logit method, GDR, and Proj networks, the frequency distribution and cumulative distribution of conditional probabilities of two groups are illustrated from Table 10.2.4 and 10.2.9 for training and testing samples respectively.

Generally, the results of these four methods indicate successful separation in producing a distribution of  $f(I/B)$  and  $f(I/N)$ . However, the MDA method demonstrated a different shape to the other three methods. The results are graphed from Figure 10.2.4 to Figure 10.2.7 for training data and from Figure 10.2.8 to Figure 10.2.11 for testing data. These histograms show that the two distributions, although they are different, have some overlap for most of these alternative techniques.

**Table 10.2.2 The Frequency Distribution and Cumulative Distribution of Z Scores for MDA Method in Training Data**

Z Score	f(I/N)	f(I/B)	F(I/N)	F(I/B)
-12 and below	2	0	2/88	0/44
-10 and below	1	0	3/88	0/44
-8 and below	18	0	21/88	0/44
-6 and below	16	0	37/88	0/44
-4 and below	29	0	66/88	0/44
-2 and below	16	1	82/88	1/44
0 and below	6	4	88/88	5/44
2 and below	0	5	88/88	10/44
4 and below	0	7	88/88	17/44
6 and below	0	8	88/88	25/44
8 and below	0	3	88/88	28/44
10 and below	0	9	88/88	37/44
12 and below	0	3	88/88	40/44
14 and below	0	1	88/88	41/44
16 and below	0	2	88/88	43/44
over 16 below	0	1	88/88	44/44
Total	88	44		
Min Score	-12.73	-3.43		
Max Score	-0.43	16.12		
Mean Score	-5.80	5.80		
Standard Deviation	2.58	4.65		

**Table 10.2.3 The Frequency Distribution and Cumulative Distribution of Z Scores for MDA Method in Testing Data**

Z Score	f(I/N)	f(I/B)	F(I/N)	F(I/B)
-35 and below	1	0	1/88	0/44
-30 and below	0	0	1/88	0/44
-25 and below	0	0	1/88	0/44
-20 and below	0	0	1/88	0/44
-15 and below	4	0	5/88	0/44
-10 and below	5	0	10/88	0/44
-5 and below	39	1	49/88	1/44
0 and below	34	5	83/88	6/44
5 and below	4	22	87/88	28/44
10 and below	1	14	88/88	42/44
15 and below	0	1	88/88	43/44
20 and below	0	0	88/88	43/44
25 and below	0	0	88/88	43/44
30 and below	0	0	88/88	43/44
over 35	0	1	88/88	44/44
Total	88	44		
Min Score	-39.90	-5.49		
Max Score	5.70	34.01		
Mean Score	-5.97	4.27		
Standard	5.51	6.10		

**Table 10.2.4 The Frequency Distribution and Cumulative Distribution of Conditional Probabilities for Logit Method in Training Data**

Probability	f(I/N)	f(I/B)	F(I/N)	F(I/B)
0.1 and below	81	3	81/88	3/44
0.2 and below	4	0	85/88	3/44
0.3 and below	1	0	86/88	3/44
0.4 and below	0	0	86/88	3/44
0.5 and below	0	0	86/88	3/44
0.6 and below	2	2	88/88	5/44
0.7 and below	0	2	88/88	7/44
0.8 and below	0	1	88/88	8/44
0.9 and below	0	3	88/88	11/44
1.0 and below	0	33	88/88	44/44
Total	88	44		
Min Score	0	0.48		
Max Score	0.52	1		
Mean Score	0.03	0.87		
Standard Deviation	0.08	0.25		

**Table 10.2.5 The Frequency Distribution and Cumulative Distribution of Conditional Probabilities for Logit Method in Testing Data**

Probability	f(I/N)	f(I/B)	F(I/N)	F(I/B)
0.1 and below	81	9	81/88	9/44
0.2 and below	0	0	81/88	9/44
0.3 and below	0	0	81/88	9/44
0.4 and below	0	0	81/88	9/44
0.5 and below	0	0	81/88	9/44
0.6 and below	0	0	81/88	9/44
0.7 and below	0	0	81/88	9/44
0.8 and below	0	0	81/88	9/44
0.9 and below	0	0	81/88	9/44
1.0 and below	7	35	88/88	44/44
Total	88	44		
Min Score	0	0		
Max Score	1	1		
Mean Score	0.08	0.79		
Standard	0.27	0.41		

**Table 10.2.6 The Frequency Distribution and Cumulative Distribution of Predicted Values for GDR Method in Training Data**

Predicted	f(I/N)	f(I/B)	F(I/N)	F(I/B)
0.1 and below	88	1	88/88	1/44
0.2 and below	0	0	88/88	1/44
0.3 and below	0	0	88/88	1/44
0.4 and below	0	0	88/88	1/44
0.5 and below	0	0	88/88	1/44
0.6 and below	0	0	88/88	1/44
0.7 and below	0	0	88/88	1/44
0.8 and below	0	0	88/88	1/44
0.9 and below	0	0	88/88	1/44
1.0 and below	0	43	88/88	44/44
Total	88	44		
Min Score	0	0.87		
Max Score	0.08	1		
Mean Score	0	0.99		
Standard Deviation	0.08	0.02		

**Table 10.2.7 The Frequency Distribution and Cumulative Distribution of Predicted Values for GDR Method in Testing Data**

Predicted	f(I/N)	f(I/B)	F(I/N)	F(I/B)
0.1 and below	77	3	77/88	3/44
0.2 and below	0	0	77/88	3/44
0.3 and below	0	0	77/88	3/44
0.4 and below	0	0	77/88	3/44
0.5 and below	0	0	77/88	3/44
0.6 and below	0	0	77/88	3/44
0.7 and below	3	1	80/88	4/44
0.8 and below	4	6	84/88	10/44
0.9 and below	0	0	84/88	10/44
1.0 and below	4	34	88/88	44/44
Total	88	44		
Min Score	0	0		
Max Score	1	1		
Mean Score	0.1	0.89		
Standard	0.27	0.26		

**Table 10.2.8 The Frequency Distribution and Cumulative Distribution of Predicted Values for Proj Method in Training Data**

Predicted	f(I/N)	f(I/B)	F(I/N)	F(I/B)
0.1 and below	88	0	88/88	0/44
0.2 and below	0	0	88/88	0/44
0.3 and below	0	0	88/88	0/44
0.4 and below	0	0	88/88	0/44
0.5 and below	0	0	88/88	0/44
0.6 and below	0	0	88/88	0/44
0.7 and below	0	0	88/88	0/44
0.8 and below	0	0	88/88	0/44
0.9 and below	0	1	88/88	1/44
1.0 and below	0	43	88/88	44/44
Total	88	44		
Min Score	0	0.87		
Max Score	0.08	1		
Mean Score	0	0.99		
Standard Deviation	0.01	0.02		

**Table 10.2.9 The Frequency Distribution and Cumulative Distribution of Predicted Values for Proj Method in Testing Data**

Predicted	f(I/N)	f(I/B)	F(I/N)	F(I/B)
0.1 and below	76	3	76/88	3/44
0.2 and below	2	0	78/88	3/44
0.3 and below	1	0	79/88	3/44
0.4 and below	0	0	79/88	3/44
0.5 and below	0	4	79/88	7/44
0.6 and below	0	0	79/88	7/44
0.7 and below	1	1	80/88	8/44
0.8 and below	4	0	84/88	8/44
0.9 and below	2	1	86/88	9/44
1.0 and below	2	35	88/88	44/44
Total	88	44		
Min Score	0	0		
Max Score	1	1		
Mean Score	0.09	0.87		
Standard	0.25	0.29		



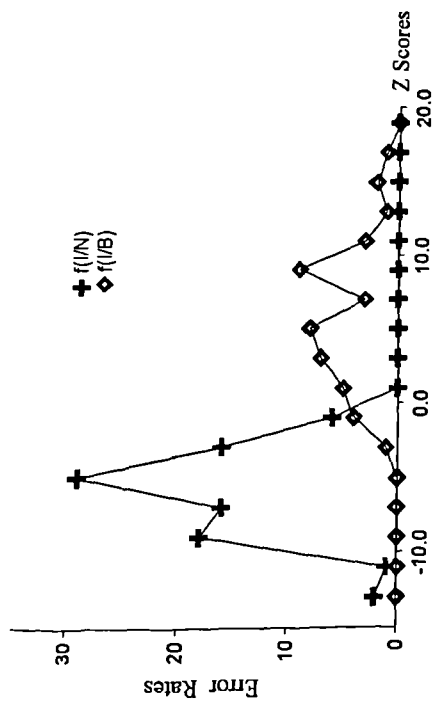


Figure 10.2.4 Frequency Distribution of Z Scores for MDA Method in Training Data

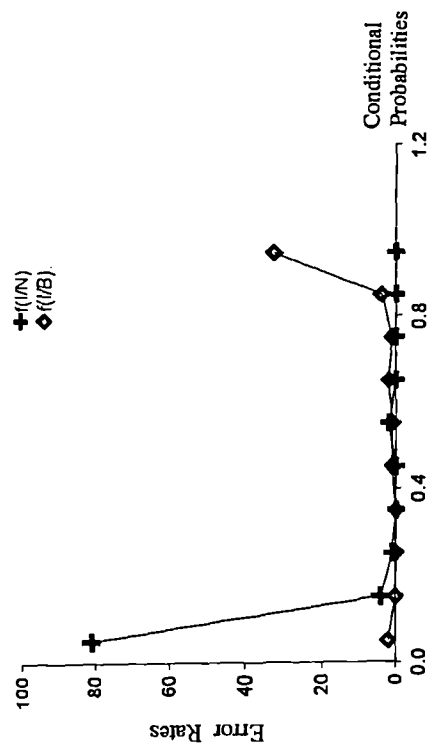


Figure 10.2.5 Frequency Distribution of Conditional Probabilities for Logit Method in Training Data

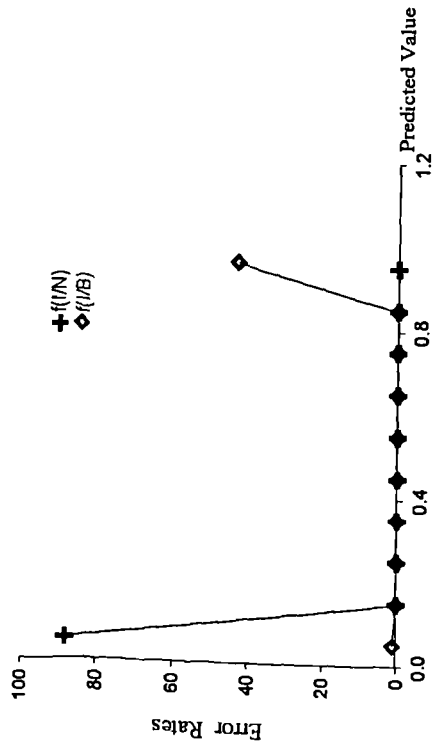


Figure 10.2.6 Frequency Distribution of Predicted Values for GDR Method in Training Data

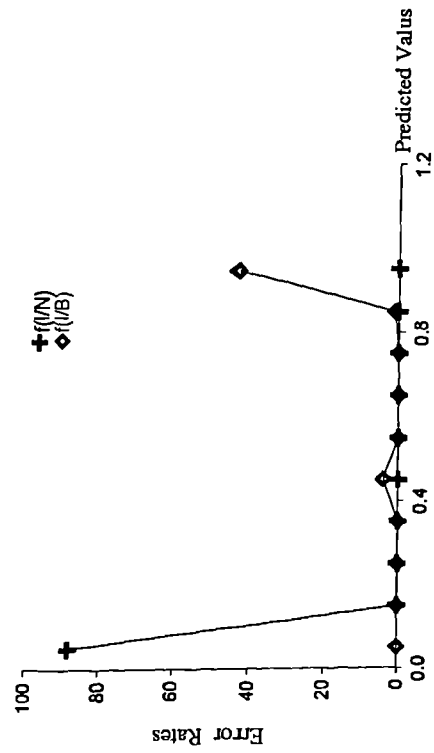


Figure 10.2.7 Frequency Distribution of Predicted Values for Proj Method in Training Data

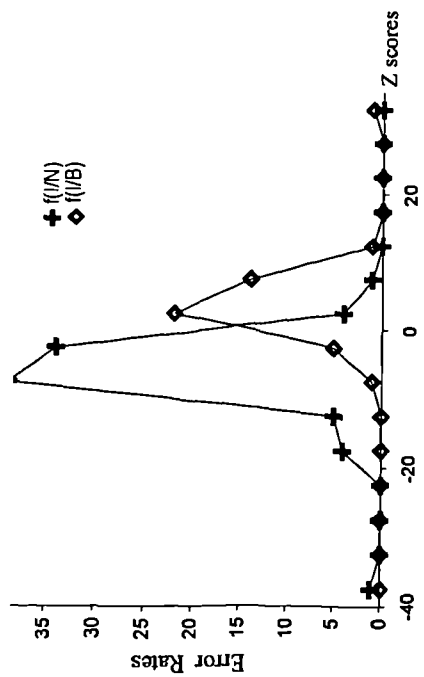


Figure 10.2.8 Frequency Distribution of Z Scores for MDA Method in Training Data

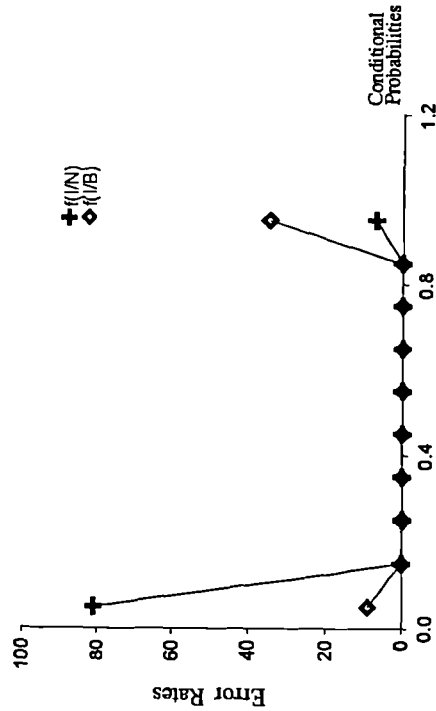


Figure 10.2.9 Frequency Distribution of Conditional Probabilities for Logit Method in Testing Data

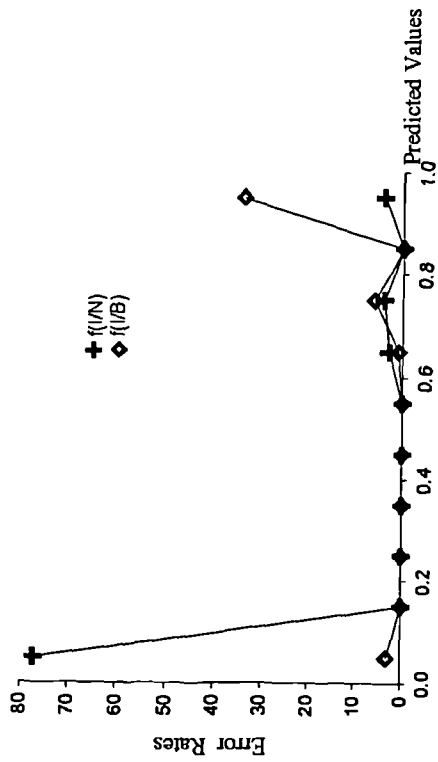


Figure 10.2.10 Frequency Distribution of Predicted Values for GDR Method in Testing Data

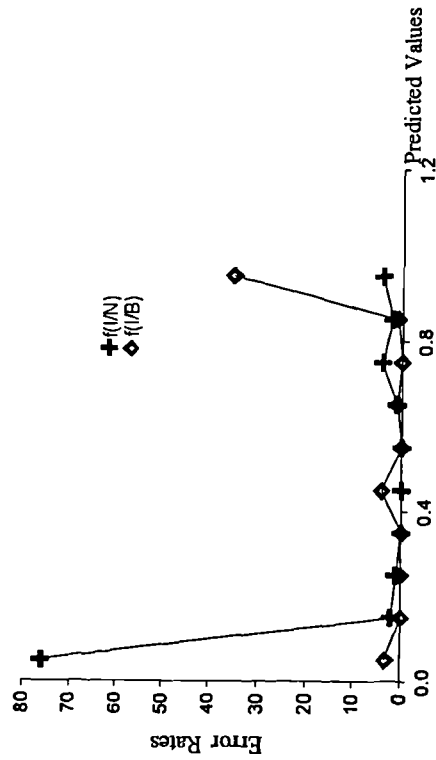


Figure 10.2.11 Frequency Distribution of Predicted Values for Proj Method in Testing Data

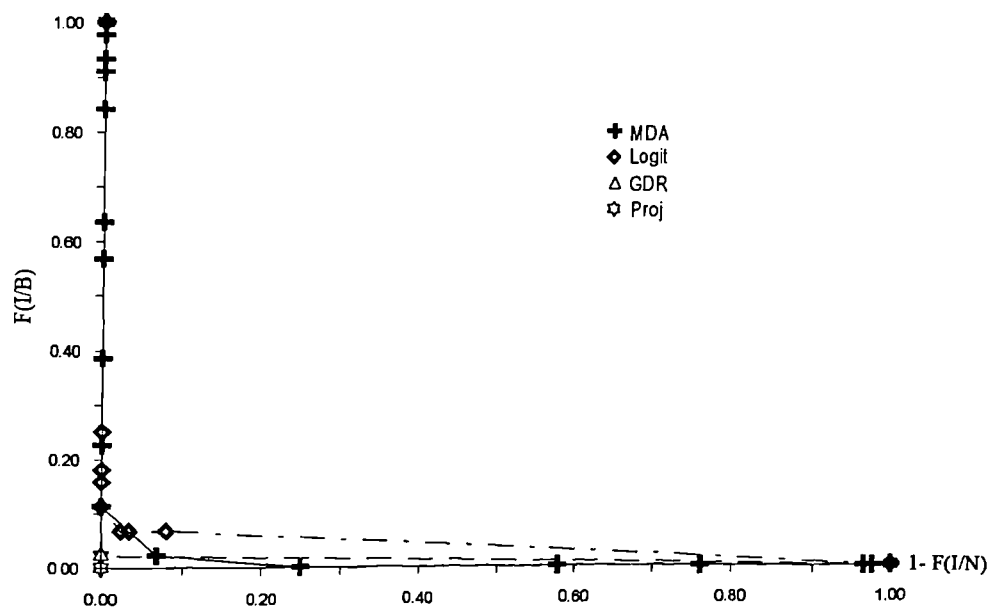


Figure 10.2.12 Comparisons on OC Curve among Four Methods for Training Data

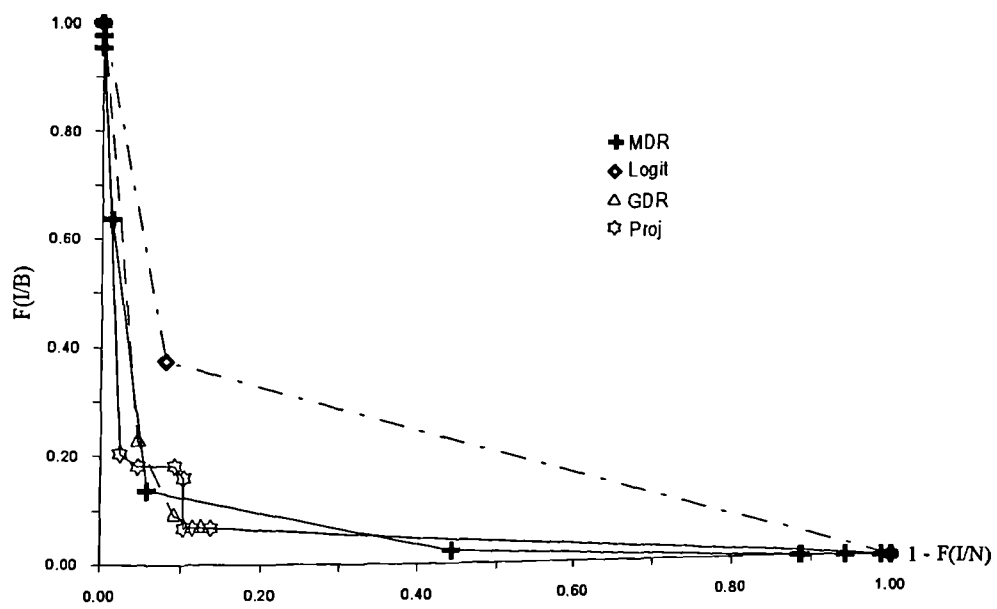


Figure 10.2.13 Comparisons on OC Curve among Four Methods for Testing Data

Up to now, we have still not been able to determine which method is most powerful in classification. As we have described previously, the OC curves of these four methods can help us to solve this question. Figure 10.2.12 shows the result of comparing OC curves for four methods on the training data. It reveals that the Projection method is superior in all situations. In this case the method produces a disjoint mapping, so that  $f(I/B)$  does not overlap with  $f(I/N)$ . It follows that the Operating Characteristic curve is the axes of diagram from (0,1) to the origin (0,0), and then to (1,0). For the other three cases, the GDR method is seen to be better than the Logit method, since over the entire range of prior probability and misclassification costs, the GDR OC curve is nearer to the axes than the Logit approach. However, the MDA and GDR do not appear to dominate each other in the training phase. The analysis indicates that in terms of learning ability, the Projection approach has the highest classification power. There are no significant differences between MDA and GDR; but the worst predictive ability is that of the Logit method.

As to generalisation ability, Figure 10.2.13 indicates that the Logit method still has the worst predictive ability in all situations. Over the entire range of the OC curve, it shows a further distance to the axes than the other three approaches, and this produces both higher Type I and Type II errors regardless of prior probabilities or error costs. For MDA, GDR, and Projection techniques, we can not conclude which is superior. The superiority depends on the subjective assessment of the situation.

The results in this empirical study are, on average, consistent with those in the simulation study. The ANNs are shown to perform at least as well as the conventional statistical methods, which have been the dominant methods in bankruptcy prediction until now. This evidence implies that ANNs indeed are promising discriminating tools in predicting business failure. However, they may suffer from the overfitting problem since they have the potential to pay undue attention to irrelevant noise in the learning phase so as to fail to identify key features and thus to lose generalisation ability.

### **10.2.2 Comparison on the Relative Importance of Predictor Variables**

Evaluating the importance of individual predictor variables is usually one of the main interests of researchers. It provides precious information to investors, and managers for an understanding of the relationship between the financial ratios and the firm's performance.

In terms of MDA method, standardised coefficients are selected to assess the relative importance of individual variables. This approach is expressed as the discriminant coefficients of each independent variable divided by its standard deviation. The standard deviation was obtained from the pooled variance-covariance matrix. The larger the standard discriminant coefficient value, the greater its contribution. The pooled variance-covariance matrix is presented in Table 10.2.10, and the relative contribution of each independent variable is then reported in Table 10.2.11.

For the Logit method, tests of the significance of individual variable coefficients can be established by Wald Chi-square statistics. The test is parallel to the discriminant analysis procedure in that coefficients are divided by their standard errors in order to measure the test statistic. The results are displayed in Table 10.2.12. The P-values for the nonparametric tests were also computed. In addition, overall tests for assessing model fit were performed. The test statistic used to assess overall fit is the -2 times Log Likelihood Ratio. This measure tests the null hypothesis that the financial ratios have no impact on the prediction of bankrupt or nonbankrupt firms. It is surprisingly observed that all predictor variables are insignificant at the 0.05 level. However, the test of overall model fitting clearly indicates that it is significant at the 0.05 or 0.01 level. Thus, the null hypothesis that the input variables have no explanatory power can be rejected. This contradictory conclusion is a result of the fact that the independent variables are correlated so that the combination of univariate variable with low individual predictive ability can generate multivariate high predictive power.

In terms of ANNs, the interpretation method suggested by Yoon et al. [1993] to extract the relative strength between each input and output unit is described as follows

$$RS_{ji} = \frac{\sum (W_{ki} * U_{jk})}{\sum_{k=0}^m ABS [\sum (W_k * U_{jk})]}$$

where

$RS_{ji}$  = the relative strength between the  $i$ th input and  $j$ th output variables

$W_{ki}$  = the weight between the  $k$ th hidden unit and  $i$ th input unit

$U_{jk}$  = the weight between the  $j$ th output and the  $k$ th hidden unit

ABS = the sign of the absolute value

The denominator measures the total strength between all of the input and output variables.

The absolute value is used because the positive strengths should not cancel out the negative

strengths. The numerator measures the strength between the *i*th input variable and the *j*th output variable and can be either positive or negative.

**Table 10.2.10 The Pooled Variance-Covariance Matrix in MDA**

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12
R1	159.83											
R2	0.58	0.03										
R3	1.96	-0.01	0.42									
R4	-57.1	-3.04	2.13	355.31								
R5	172.4	0.89	1.82	-87.78	219.32							
R6	1.48	0.05	-0.04	-5.08	1.62	0.24						
R7	-68.1	-3.02	1.95	354.61	-97.08	-5.6	392.24					
R8	3.17	0.08	-0.05	-9.01	4.01	0.25	-10.22	0.55				
R9	-0.28	0.01	-0.03	-0.28	-0.2	0.04	-0.35	0.03	0.02			
R10	39.29	1.32	2.51	-120	55.33	1.01	-146.6	3.94	0.1	140.2		
R11	-0.83	0	-0.06	0.17	-0.74	0.05	-0.2	0.01	0.02	-0.48	0.04	
R12	-0.01	0	0	-0.02	-0.028	0	-0.03	0	0	0.01	0	0.001

**Table 10.2.11 Relative Contribution Tests of Each Independent Variables and Its Rank for MDA Method**

Ratios	Coefficient	Standard Deviation	Standardised Coefficient	Rank
R1	-5.0852	12.6425	-0.0131	9
R2	-0.6438	0.1761	-14.5847	4
R3	-2.5679	0.6481	-5.9579	5
R4	-3.8612	18.8496	-0.0002	11
R5	0.0298	14.8118	-0.0055	8
R6	-0.0821	0.494	4.9833	6
R7	-0.2138	19.8049	0.0108	10
R8	1.3167	0.7396	-1.7803	7
R9	3.2973	0.1378	23.9211	3
R10	-0.0212	33.7668	-0.0006	12
R11	7.8705	0.1897	-41.4813	2
R12	-36.491	0.0316	1,153.95	1

**Table 10.2.12 Test of Individual Coefficients and Model Fitting  
for Logit Approach**

Ratio	Parameter	Estimated	Standardised	Wald	P value	Rank
R1	0.2127	516.1	1.5138	0	0.9997	12
R2	145	49,571.1	14.04	0	0.9977	4
R3	-25.7782	9,841.2	-9.2066	0	0.9979	7
R4	0.5999	1,164.2	6.2347	0	0.9996	8
R5	-4.2589	382.3	-34.7729	0.0001	0.9911	1
R6	-53.0101	14,287	-14.4474	0	0.997	3
R7	1.0269	724	11.2128	0	0.9989	6
R8	-29.6826	6,564.7	-12.1076	0	0.9964	5
R9	37.7197	72,537.3	2.8411	0	0.9996	10
R10	-0.3118	66.6016	-5.8044	0	0.9963	9
R11	163.5	42,750.6	17.0566	0	0.9969	2
R12	157.8	185,326	1.5562	0	0.9993	11
Criteria for Assessing Model Fit						
				$\chi^2$ for Covariate	p value	
-2 Times log likelihood ratio				168.4	0.1	

ABS = the sign of the absolute value

For the GDR and Projection methods, the weights from the input layer to the hidden layer and from the hidden layer to the output layer are indicated in Table 10.2.13 and 10.2.14 respectively. The interpretations of the relative strength between an input variable and an output variable are then shown in Table 10.2.15. For ease of comparison the results in MDA and Logit methods are repeated here.

It can be seen that the ranking of the relative contributions of the twelve predictor variables for these alternative methods seems to be quite different. Thus, we performed a correlation analysis for the ranking to understand if there exists any relationship between different techniques for these rankings. The nonparametric Spearman's Rank-Order Correlation analysis was chosen to measure this relationship. The result is shown in Table 10.2.16.

From Table 10.2.16 we notice that the all Spearman correlation coefficients are very low, and that there are even many negative signs between different methods. For any two methods there has been shown to be no significant relationship in the ranking of the relative contributions of explanatory variables at 0.05 level. This inconsistent but interesting outcome is believed to be worthy of further study.

**Table 10.2.13 The Neural Network Weights for GDR Method**

(a) From input layer to hidden layer

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12
Node	-39.98	11.57	-26.28	4.25	-50.75	-9.68	27.58	-27.17	7.33	-11.49	24.01	7.21
Node	-16.87	3.28	-9.77	4.07	-20.95	-3.6	13.38	-11.02	4.76	-4.95	11.03	3.86
Node	10.54	-3.14	6.98	-0.69	12.89	2.74	-6.34	7.24	-1.02	3.05	-5.35	-1.56
Node	-22.42	8.01	-13.93	2.57	-28.76	-4.99	15.78	-14.46	3.96	-6.48	13.76	4.3
Node	37.88	-14.62	24.34	-3.16	47.51	7.62	-25.48	23.2	-8.63	10.41	-23.98	-8.32
Node	-38.61	14.41	-24.26	2.97	-50	-8.81	26.27	-26.4	6.79	-11.14	23.47	6.9
Node	-21.19	8.11	-12.18	2.46	-26.98	-2.88	15.19	-12.84	5.52	-6.1	14.55	4.21
Node	30.91	-10.32	19.88	-2.49	38.97	8.1	-20.38	21.59	-4.37	8.78	-17.32	-4.88
Node	30.74	-8.42	20.77	-3.46	38.29	7.96	-20.63	20.69	-5.03	8.5	-17.16	-5.95

(b) From hidden layer to output layer

	Node 1	Node 2	Node 3	Node 4	Node 5	Node 6	Node 7	Node 8	Node 9
Output	-0.06	-0.17	0.02	-0.21	0.23	-0.1	0.03	0.01	-0.19

**Table 10.2.14 The Neural Network Weights for Proj Method**

(a) From input layer to hidden layer

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12
Node	-39.98	11.57	-26.28	4.25	-50.75	-9.68	27.58	-27.17	7.33	-11.49	24.01	7.21
Node	-16.87	3.28	-9.77	4.07	-20.95	-3.6	13.38	-11.02	4.76	-4.95	11.03	3.86
Node	10.54	-3.14	6.98	-0.69	12.89	2.74	-6.34	7.24	-1.02	3.05	-5.35	-1.56
Node	-22.42	8.01	-13.93	2.57	-28.76	-4.99	15.78	-14.46	3.96	-6.48	13.76	4.3
Node	37.88	-14.62	24.34	-3.16	47.51	7.62	-25.48	23.2	-8.63	10.41	-23.98	-8.32
Node	-38.61	14.41	-24.26	2.97	-50	-8.81	26.27	-26.4	6.79	-11.14	23.47	6.9
Node	-21.19	8.11	-12.18	2.46	-26.98	-2.88	15.19	-12.84	5.52	-6.1	14.55	4.21
Node	30.91	-10.32	19.88	-2.49	38.97	8.1	-20.38	21.59	-4.37	8.78	-17.32	-4.88
Node	30.74	-8.42	20.77	-3.46	38.29	7.96	-20.63	20.69	-5.03	8.5	-17.16	-5.95

(b) From hidden layer to output layer

	Node 1	Node 2	Node 3	Node 4	Node 5	Node 6	Node 7	Node 8	Node 9
Output	-48.55	-48.71	-48.92	-49.76	-49.5	-47.819	-48.964	-48.497	-48.77



**Table 10.2.15 Relative Strengths between Each Input Variables and Output  
for Both GDR and Proj Methods  
and the Rank of the Input Variables for Four Methods**

Ratio	Relative Strength	Relative Strength	Rank for GDR	Rank for Proj	Rank for MDA	Rank for Logit
R1	0.5654	-0.1409	2	1	9	12
R2	-0.2065	-0.0471	7	9	4	4
R3	0.3481	-0.1209	6	3	5	7
R4	-0.062	-0.0583	12	6	11	8
R5	0.7203	-0.1262	1	2	8	1
R6	0.1163	-0.0443	11	10	6	3
R7	-0.3982	-0.0523	3	8	10	6
R8	0.3599	-0.0037	5	7	7	5
R9	-0.1257	0.0058	9	12	3	10
R10	0.1619	-0.0995	8	4	12	9
R11	-0.363	-0.0126	4	11	2	2
R12	-0.1165	-0.0605	10	5	1	11

**Table 10.2.16 Spearman Correlation Analysis for Ranks  
of Relative Importance of Predictor Variables among Four Methods**

Spearman Correlation Coefficients / Prob > |R| under Ho: Rho=0

	MDA	LOGIT	GDR	PROJ
MDA	—			
LOGIT	0.09790 0.7621	—		
GDR	-0.13287 0.6806	0.27972 0.3786	—	
PROJ	-0.45455 0.1377	-0.29371 0.3541	0.37063 0.2356	—

### 10.3 The Impact of Sample Size on Predictive Ability

The second comparison in the empirical study is to assess whether the sample size affects predictive ability in terms of learning and generalisation for each of the four discriminating methods. For certain levels of sample size we also test if there is a significant preference for any technique. If the predictive ability of any particular method is superior to others and insensitive to the sample size, it can solve the difficulties of obtaining a large sample size, which is often not available. The relevant hypotheses in this section are stated as follows

$H_{20}$  : The rate of misclassification for each of the four discriminating techniques is not affected by the sample size.

$H_{21}$  : There is no significant difference in predictive performance among the four alternative techniques for different levels of sample size.

$H_{22}$  : The neural networks are not more robust than the statistical discriminating methods to sample size in predictive performance.

#### 10.3.1 The Results and Analyses

This experiment was evaluated in equal proportions of failing to nonfailing samples. Small, medium, and large of samples of 30, 60, and 120 observations were replicated 20 times. Tables 10.3.1 and 10.3.2 report the average Type I, Type II and Overall error rates at different sample size levels for four discriminating methods in training and testing data respectively.

First we examine the impact of sample size on each method. Secondly we compare the classification accuracy of the four techniques. Figures 10.3.1 to 10.3.8 display the results on training and testing data through graphic analysis for each of four approaches. Three points may be summarised from these results:

- (1) In the light of the training sample, it seems that, for all methods, the misclassification rates do not take advantage of large sample size. On the contrary, the classification accuracy decreases when the sample size increases. However, for the testing sample, the error rates are shown to decrease as the sample size

Table 10.3.1 Average Misclassification Rates vs. Sample Size  
for Four Methods in Training Sample

	Type I Error		
	30	60	120
MDA	4.000 (4.537)	6.000(3.684)	8.584(3.212)
Logit	0.000(0.000)	2.835(6.9515)	6.750(5.933)
GDR	4.336(3.264)	6.501(2.753)	6.669(4.129)
Proj	3.001(4.033)	2.999(2.395)	4.583(2.410)
	Type II Error		
	30	60	120
MDA	2.668(5.027)	1.666(2.023)	4.416(2.552)
Logit	0.000(0.000)	3.000(7.404)	5.965(5.337)
GDR	2.001(3.809)	2.999(2.841)	6.085(4.129)
Proj	0.334(1.491)	2.999(2.841)	3.500(1.700)
	Overall Error		
	30	60	120
MDA	3.34(3.587)	3.833(1.957)	6.50(2.172)
Logit	0.000(0.000)	2.9175(7.152).	6.358(5.605)
GDR	3.168(2.758)	4.750(2.312)	6.377(2.482)
Proj	1.667(2.295)	2.999(1.920)	4.412(1.628)

\* The numbers in the table are indicated in percentage

\* The numbers within parentheses are the standard deviations of the misclassified

Table 10.3.2 Average Misclassification Rates vs. Sample Size  
for Four Methods in Testing Sample

	Type I Error		
	30	60	120
MDA	17.999(12.815)	12.834(5.950)	11.751(3.220)
Logit	16.334(9.041)	9.833(5.012)	6.333(2.268)
GDR	11.000(8.726)	9.001(5.309)	10.167(2.588)
Proj	10.999(7.262)	9.168(5.173)	5.583(2.182)
	Type II Error		
	30	60	120
MDA	12.000(6.701)	4.500(4.227)	6.250(2.957)
Logit	16.665(11.342)	8.999(7.807)	6.750(3.991)
GDR	7.000(8.508)	6.834(5.565)	6.417(4.234)
Proj	6.334(5.913)	4.975(4.249)	6.333(3.226)
	Overall Error		
	30	60	120
MDA	14.999(7.375)	8.667(3.846)	9.000(2.305)
Logit	16.500(6.708)	9.416(4.336)	6.542 (2.046)
GDR	9.000(5.197)	7.918(3.746)	8.292(2.573)
Proj	8.667(4.379)	7.071(2.619)	5.958(1.488)

increases with all techniques. It indicates that during the learning process, although a small sample size is favourable, nevertheless, it may overfit the data, overemphasising on the irrelevant details and noise, ignoring the main pattern in the data, so that its capacity to generalise in prediction deteriorates. Therefore, a small sample size may not be recommended in the classification problem if generalisation ability is our primary objective.

- (2) Regardless of whether ANNs or STMs are used, with a balanced proportion of two groups (i.e., 1:1 base rate) in this experiment, the Type I error rates are always higher than the Type II error rates for training and for almost testing data (except for the sample size 30 in the Logit approach) across all small, medium and large sample size situations. In another words, the probability of misclassifying the bankrupt firms as nonbankrupt firms is much greater than that of misclassifying nonbankruptcy as bankruptcy. The results imply that the way to classify firms into groups on the basis of this real financial data for the four methods is not inconsistent, and further reinforce the earlier findings in our simulation study that ANNs have a tendency to reducing Type II error rates rather than Type I error rates.
- (3) With respect to classification accuracy in the testing data, it is observed that when the sample size increases from a medium (60) to large (120) data set, the predictive ability has not necessarily improved accordingly. This may be the result of the law of diminishing marginal returns. Moreover, Freed and Glover [1986] have pointed out that a significantly large sample size may complicate the task of dissection and evaluation. Hence, too large a sample size does not seem necessary in the prediction of bankruptcy.

After examining the outcome of misclassification rates versus sample sizes in each method, we will now compare the three types of errors for the four methods. Figures 10.3.9 to 10.3.11 illustrate the Type I, Type II and Overall error rates against different levels of sample size individually for the four methods on the basis of training data. Figures 10.3.12 to 10.3.14 present the results on testing data. There are four findings for these comparisons

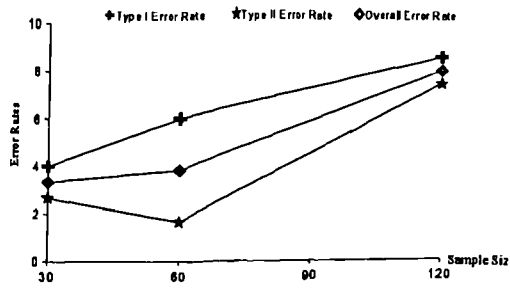


Figure 10.3.1 Error Rates vs. Sample Size in MDA for Training Data

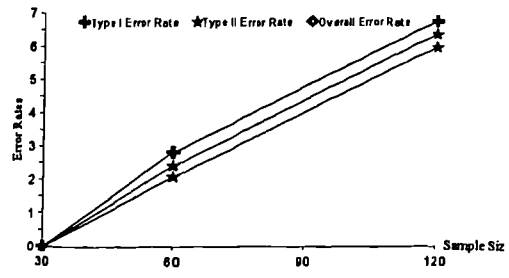


Figure 10.3.3 Error Rates vs. Sample Size in Logit for Training Data

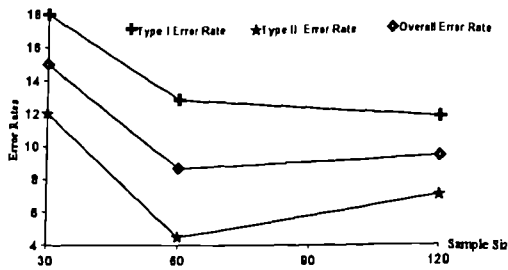


Figure 10.3.2 Error Rates vs. Sample Size in MDA for Testing Data

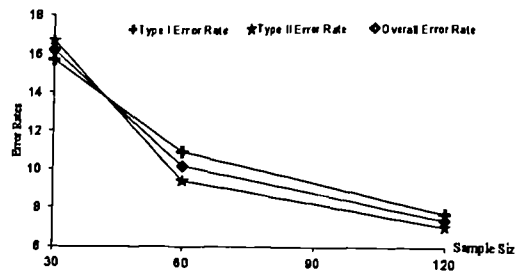


Figure 10.3.4 Error Rates vs. Sample Size in Logit for Testing Data

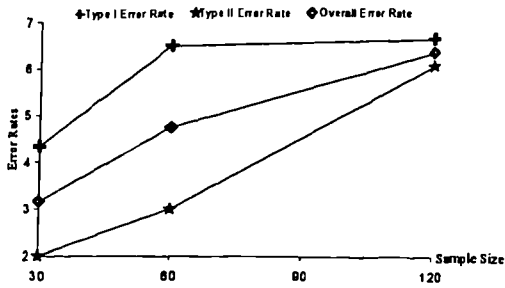


Figure 10.3.5 Error Rates vs. Sample Size in GDR for Training Data

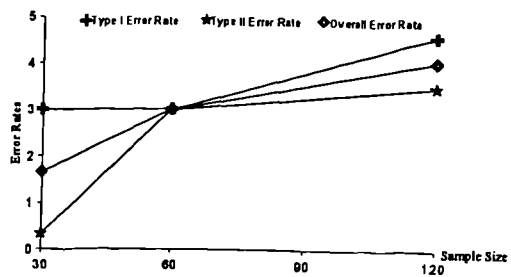


Figure 10.3.7 Error Rates vs. Sample Size in Proj for Training Data

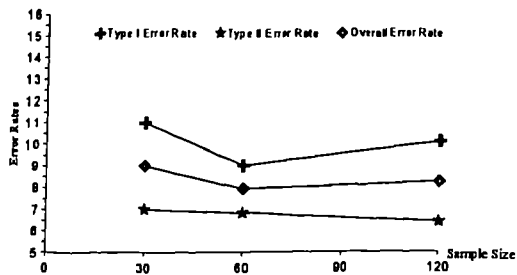


Figure 10.3.6 Error Rates vs. Sample Size in GDR for Testing Data

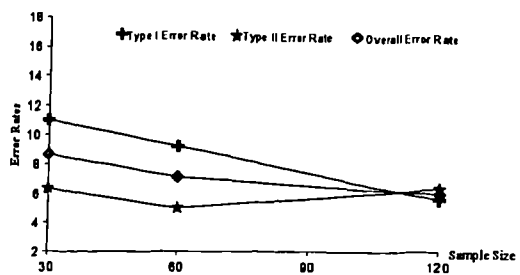


Figure 10.3.8 Error Rates vs. Sample Size in Proj for Testing Data

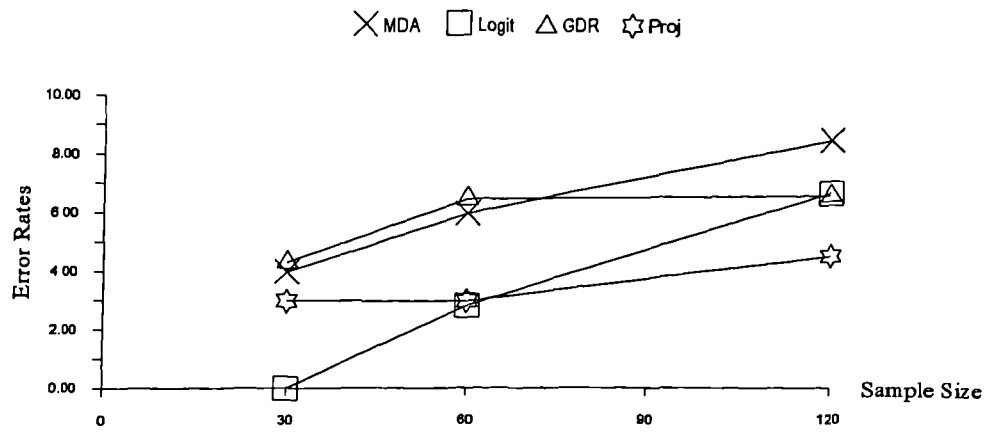


Figure 10.3.9 Type I Error vs. Sample Size for Four Methods for Training Data

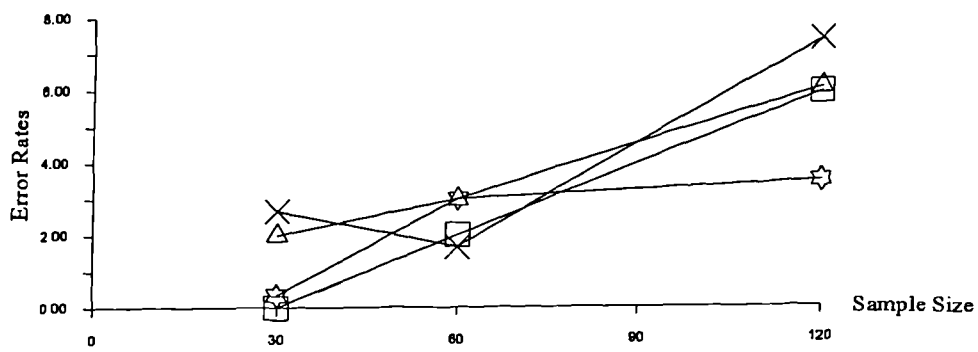


Figure 10.3.10 Type II Error vs. Sample Size for Four Methods for Training Data

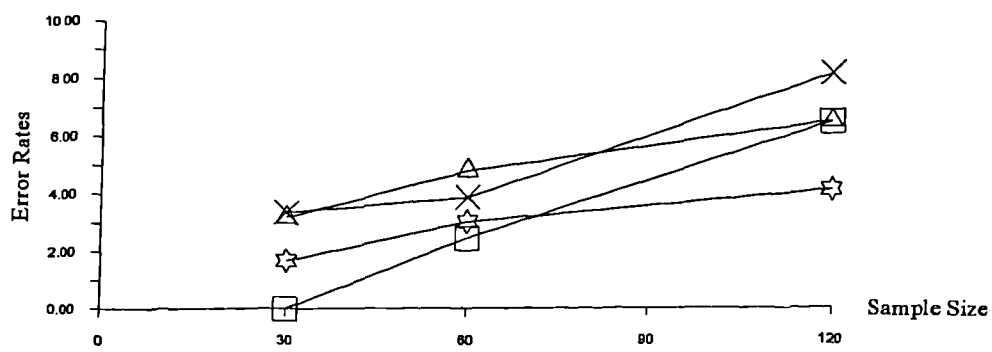


Figure 10.3.11 Overall Error vs. Sample Size for Four Methods for Training Data

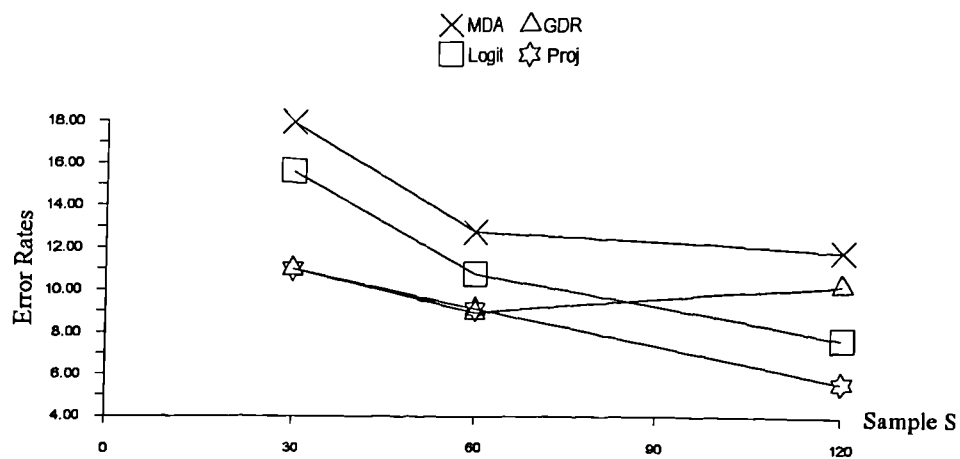


Figure 10.3.12 Type I Error vs. Sample Size for Four Methods for Testing Data

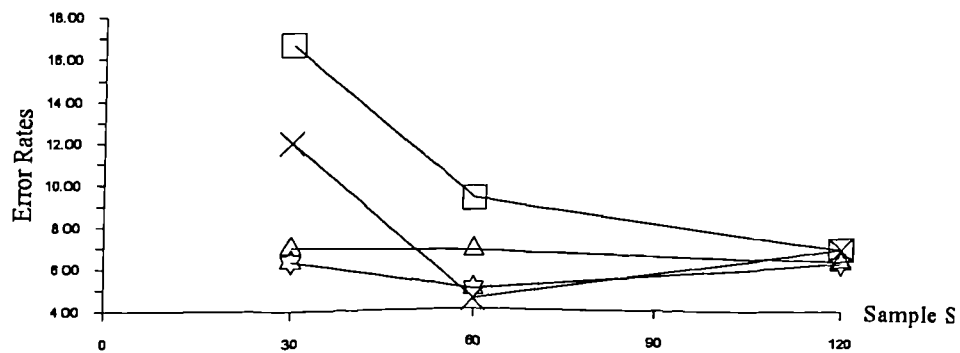


Figure 10.3.13 Type II Error vs. Sample Size for Four Methods for Testing Data

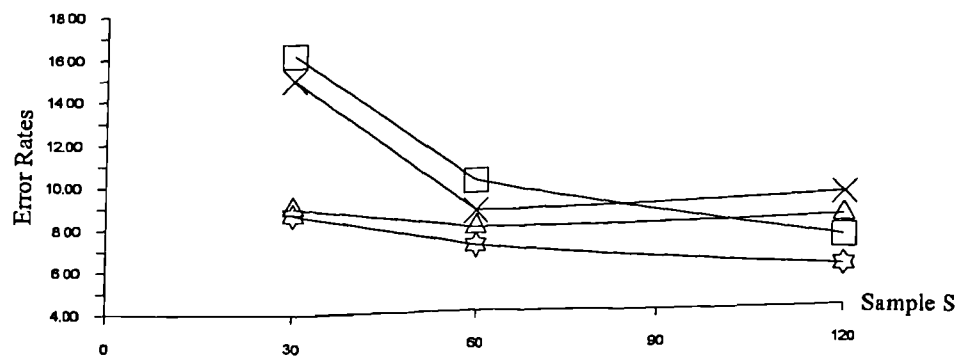


Figure 10.3.14 Overall Error vs. Sample Size for Four Methods for Testing Data

- (1) Observing the misclassification rates for the training data, we find that, in addition to the Projection method's showing lower Type I error rates and thus lower overall error rates than MDA, no method produces overwhelming superiority across all sample sizes. Thus, for learning methodology, we can not conclude that neural networks are better than the statistical techniques or vice versa.
- (2) For testing data, most cases indicate that whatever the type of error, the neural networks offer better predictive abilities than statistical methods. Nevertheless, this advantage decreases as the sample size increases.
- (3) Judging from the results of standard deviation in the three kinds of error rates, we may suggest that the ANNs generate more stable classification performance than do STMs. However, this stability becomes worse in a small size of sample relative to a large size of sample, which conforms to other studies' suggestions. Thus, small sample size should be avoided if possible when implementing ANNs.
- (4) Generally speaking, the ANNs are more robust to sample size than the statistical approaches. That is, the change of classification performance is smaller in both GDR and Proj than MDA & Logit as the sample size changes.

#### **10.4 The Bias of Choice-Based Sample Design and Its Elimination by Applying WCOP Procedure**

As we have mentioned before, for most bankruptcy prediction models, the samples were drawn based on the knowledge of the dependent variables (i.e., bankruptcy or nonbankruptcy) instead of through an exogenous random sampling design. This violation of the random sampling assumption can lead to a choice-based sample bias of probability estimate. This bias can decrease as the proportion of the two groups in the sample approaches the proportion in the population [Zmijewski, 1984]. However, this implies that very large random samples are needed in order to obtain information on the rare occurrences of bankruptcy if the effects of choice-based sample bias are to be minimised. Consequently, for the matched sampling design, which attempts to eliminate the influences of other factors such as size or industry, some adjusted techniques should be used in order to avoid the choice-based sample bias.



In this section we use six choice-based samples with different proportions of bankrupt and nonbankrupt firms in order to demonstrate the existence of a choice-based sample bias as well as the elimination of this bias when the adjusted WCOP (weighted cutoff point procedure) approach is applied. The proportions of bankruptcy to nonbankruptcy in the six estimation samples are 15:15, 15:30, 15:60, 15:75, 15:120, and 15:150. The resulting bankrupt firm frequency rates are 0.5, 0.333, 0.2, 0.167, 0.11, and 0.09 respectively.

The choice-based sampling issue is examined by using both unweighted assessment and the WCOP model shown in equation (8.3.6), which was proposed in Chapter Eight, on six choice-based estimation samples. The samples have decreasing bankrupt firm frequency rates so that the bias induced by estimating the model via an unadjusted procedure can be assessed as the sample selection probabilities approach the population probability.

The relevant hypotheses in this experiment are stated as follows

$H_{23}$  : The Type I, Type II and Overall error rates have no functional relationship with the decreasing choice-based sample frequency rate in each method when using an unadjusted procedure.

$H_{24}$  : If the choice-based sample bias exists, it does not decrease when the proportion of two groups in the sample approaches the prior probability in the population.

$H_{25}$  : The Type I, Type II and Overall error rates have no functional relationship with the decreasing choice-based sample frequency rate in each method when using the WCOP procedure.

#### **10.4.1 The Results and Analyses**

The above two hypotheses  $H_{23}$  and  $H_{25}$  are equivalent to testing whether there exists a choice-based sample bias or not when using or not using an adjusted process. The Pearson correlation coefficients between the sample frequency rate and the group error rates are used to indicate the existence of a choice-based sample bias. Correlation coefficients consistent with this bias would be negative for failing firms and positive for nonfailing firms. Put another way, higher bankrupt firm sample frequency rates cause lower Type I estimated error rates. However, a higher bankrupt firm proportion in samples causes

higher Type II error rates. This bias would thus be an increasing function of the difference between the sample selection probability and the population probability

Conversely, if a bias does not exist, then there should be no relation between the bankrupt firm sample frequency rate of the choice-based estimation samples and the various results of group classification and prediction error rates.

Tables 10.4.1 to 10.4.8 report the results of Type I, Type II, Overall errors and Pearson correlation coefficients for MDA, Logit, GDR and Proj methods across the six cases. The presence of a bias is examined by comparing the unweighted and weighted estimations in panel A and panel B respectively.

In training data, the results generally indicate the existence of a bias and the overclassification of bankrupt firms when using an unweighted process. The correlation coefficients between the various proportions of two groups in the sample and the percentage of bankrupt firms misclassified in panel A are shown to be significantly higher than the results using the WCOP procedure in panel B. For instance, the unadjusted correlation coefficients are -0.906, -0.660, -0.905 and -0.905 for MDA, Logit, GDR and Proj respectively, which are a contrast to -0.578, -0.389, -0.700, and -0.680 of correlation coefficients based on the WCOP results.

For testing data, the outcome of unweighted procedure still exhibits the overclassification bias in the bankrupt group and the underclassification bias in the nonbankrupt group. Furthermore, the overall misclassification correlation is positive, indicating the correct prediction increase when the samples which are less biased are used. Meanwhile, all results of WCOP adjustment have relative lower correlations compared with those of unweighted procedure, providing equivalent conclusions to those drawn from the training data.

The tests reported in this section empirically demonstrate the effects of choice-based samples on classification accuracy. The evidence indicates that the null hypothesis  $H_{23}$  should be rejected since the functional relationship of misclassification rates to the differences between the sample selection probability and the population probability is presented using unadjusted estimation techniques. However, this bias decreases when the sample frequency rate of bankruptcy approaches the prior probability of bankruptcy in the population. Therefore, the null hypothesis  $H_{24}$  should be rejected as well. On the other hand, the test also demonstrate how using the WCOP process to estimate such models on choice-based samples eliminates most, if not all, of the bias.

**Table 10.4.1 Comparison of Unweighted and WCOP on Classification Accuracy Across Alternative Estimation Samples in MDA Using Training Sample Results**

	Choice-Based Estimation Sample <sup>1</sup>						Pearson
	15:15	15:30	15:60	15:75	15:120	15:150	Correlation
	(0.5)	(0.333)	(0.2)	(0.167)	(0.11)	(0.09)	Coefficients <sup>2</sup>
<b>Panel A- Unweighted Results</b>							
Training Data <sup>3</sup>							
Bankrupt (Type I Error)	0	6.67	6.67	6.67	13.33	13.33	-0.906
Nonbankrupt (Type II Error)	6.67	3.33	3.33	2.67	1.67	2.67	0.918
Overall	3.33	4.44	4	3.34	2.97	3.64	0.211
<b>Panel B- Weighted Results</b>							
Training Data							
Bankrupt (Type I Error)	6.67	13.33	6.67	13.33	13.33	13.33	-0.578
Nonbankrupt (Type II Error)	0	0	1.67	1.33	1.67	2.67	-0.871
Overall	3.33	4.44	2.67	3.33	2.97	3.64	0.278

**Table 10.4.2 Comparison of Unweighted and WCOP on Classification Accuracy Across Alternative Estimation Samples in MDA Using Testing Sample Results**

	Choice-Based Estimation Sample <sup>1</sup>						Pearson
	15:15	15:30	15:60	15:75	15:120	15:150	Correlation
	(0.5)	(0.333)	(0.2)	(0.167)	(0.11)	(0.09)	Coefficients <sup>2</sup>
<b>Panel A- Unweighted Results</b>							
Training Data <sup>3</sup>							
Bankrupt (Type I Error)	6.67	13.33	13.33	13.33	20	20	-0.905
Nonbankrupt (Type II Error)	13.33	6.67	6.67	4	2.5	2	0.963
Overall	10	8.89	7.99	5.56	4.44	3.64	0.919
<b>Panel B- Weighted Results</b>							
Training Data							
Bankrupt (Type I Error)	13.33	20	20	13.33	20	20	-0.496
Nonbankrupt (Type II Error)	13.33	0	5	4	1.67	2	0.708
Overall	13.15	6.67	8	5.56	3.71	3.64	0.922

<sup>1</sup> Number of bankrupt : number of nonbankrupt firms in the choice-based estimation sample.

The bankrupt firms sample frequency rate(number of bankrupt firms/total number of sample firms)

<sup>2</sup> Pearson correlation coefficients between the estimation sample frequency rate and result reported in the corresponding row

<sup>3</sup> Percentage of firms misclassified

**Table 10.4.3 Comparison of Unweighted and WCOP on Classification Accuracy Across Alternative Estimation Samples in Logit Using Training Sample Results**

	Choice-Based Estimation Sample <sup>1</sup>						Pearson
	15:15	15:30	15:60	15:75	15:120	15:150	Correlation
	(0.5)	(0.333)	(0.2)	(0.167)	(0.11)	(0.09)	Coefficients <sup>2</sup>
<b>Panel A- Unweighted Results</b>							
Training Data <sup>3</sup>							
Bankrupt (Type I Error)	0	0	0	0	6.67	6.67	-0.66
Nonbankrupt (Type II Error)	0	0	0	0	0	0	NA
Overall	0	0	0	0	0.74	0.61	-0.651
<b>Panel B- Weighted Results</b>							
Training Data							
Bankrupt (Type I Error)	0	6.67	0	0	6.67	6.67	-0.389
Nonbankrupt (Type II Error)	0	0	0	0	0	0	NA
Overall	0	2.22	0	0	0.74	0.61	0.063

**Table 10.4.4 Comparison of Unweighted and WCOP on Classification Accuracy Across Alternative Estimation Samples in Logit Using Testing Sample Results**

	Choice-Based Estimation Sample <sup>1</sup>						Pearson
	15:15	15:30	15:60	15:75	15:120	15:150	Correlation
	(0.5)	(0.333)	(0.2)	(0.167)	(0.11)	(0.09)	Coefficients <sup>2</sup>
<b>Panel A- Unweighted Results</b>							
Training Data <sup>3</sup>							
Bankrupt (Type I Error)	6.67	6.67	6.67	6.67	13.33	20	-0.626
Nonbankrupt (Type II Error)	13.33	6.67	8.33	5.33	0.83	0.67	0.904
Overall	10	6.67	8	5.55	2.22	2.43	0.861
<b>Panel B- Weighted Results</b>							
Training Data							
Bankrupt (Type I Error)	20	6.67	6.67	13.33	6.67	20	0.235
Nonbankrupt (Type II Error)	6.67	3.33	8.33	5.33	2.5	0.67	0.485
Overall	13.34	4.44	8	6.66	2.96	2.43	0.813

<sup>1</sup> Number of bankrupt : number of nonbankrupt firms in the choice-based estimation sample.

The bankrupt firms sample frequency rate(number of bankrupt firms/total number of sample firms)

<sup>2</sup> Pearson correlation coefficients between the estimation sample frequency rate and result reported in the corresponding row

<sup>3</sup> Percentage of firms misclassified

**Table 10.4.5 Comparison of Unweighted and WCOP on Classification Accuracy Across Alternative Estimation Samples in GDR Using Training Sample Results**

	Choice-Based Estimation Sample <sup>1</sup>						Pearson
	15:15	15:30	15:60	15:75	15:120	15:150	Correlation
	(0.5)	(0.333)	(0.2)	(0.167)	(0.11)	(0.09)	Coefficients <sup>2</sup>
<b>Panel A- Unweighted Results</b>							
Training Data <sup>3</sup>							
Bankrupt (Type I Error)	0	0	6.67	6.67	13.33	13.33	-0.905
Nonbankrupt (Type II Error)	0	0	1.67	1.33	1.33	1.33	-0.859
Overall	0	0	2.67	2	2.66	2.42	-0.894
<b>Panel B- Weighted Results</b>							
Training Data	6.67	6.67	13.33	6.67	13.33	13.33	-0.7
Bankrupt (Type I Error)	0	0	0	0	2.5	1.33	-0.604
Nonbankrupt (Type II Error)	3.33	2.22	1.334	1.11	3.7	2.42	0.232
Overall							

**Table 10.4.6 Comparison of Unweighted and WCOP on Classification Accuracy Across Alternative Estimation Samples in GDR Using Testing Sample Results**

	Choice-Based Estimation Sample <sup>1</sup>						Pearson
	15:15	15:30	15:60	15:75	15:120	15:150	Correlation
	(0.5)	(0.333)	(0.2)	(0.167)	(0.11)	(0.09)	Coefficients <sup>2</sup>
<b>Panel A- Unweighted Results</b>							
Training Data <sup>3</sup>							
Bankrupt (Type I Error)	0	6.67	6.67	13.33	20	20	-0.914
Nonbankrupt (Type II Error)	6.67	6.67	6.67	4.17	2.5	1.33	0.779
Overall	3.34	6.67	6.67	5.7	4.44	3.03	-0.021
<b>Panel B- Weighted Results</b>							
Training Data							
Bankrupt (Type I Error)	6.67	13.33	6.67	13.33	26.67	20	-0.682
Nonbankrupt (Type II Error)	0	0	3.33	4.17	1.67	1.33	0.586
Overall	3.34	4.44	4	5.7	4.45	3.03	0.217

<sup>1</sup> Number of bankrupt : number of nonbankrupt firms in the choice-based estimation sample.

The bankrupt firms sample frequency rate(number of bankrupt firms/total number of sample firms)

<sup>2</sup> Pearson correlation coefficients between the estimation sample frequency rate and result reported in the corresponding row

<sup>3</sup> Percentage of firms misclassified

**Table 10.4.7 Comparison of Unweighted and WCOP on Classification Accuracy Across Alternative Estimation Samples in Proj Using Training Sample Results**

	Choice-Based Estimation Sample <sup>1</sup>						Pearson
	15:15	15:30	15:60	15:75	15:120	15:150	Correlation
	(0.5)	(0.333)	(0.2)	(0.167)	(0.11)	(0.09)	Coefficients <sup>2</sup>
<b>Panel A- Unweighted Results</b>							
Training Data <sup>3</sup>							
Bankrupt (Type I Error)	0	0	6.67	6.67	13.33	13.33	-0.905
Nonbankrupt (Type II Error)	0	0	2.66	2	0.83	0.67	-0.495
Overall	0	0	3.46	2.78	2.22	1.82	-0.742
<b>Panel B- Weighted Results</b>							
Training Data							
Bankrupt (Type I Error)	6.67	0	13.33	13.33	13.33	13.33	-0.68
Nonbankrupt (Type II Error)	0	0	0	0	0	0.67	-0.449
Overall	3.34	0	2.67	2.67	2.22	1.82	0.099

**Table 10.4.8 Comparison of Unweighted and WCOP on Classification Accuracy Across Alternative Estimation Samples in Proj Using Testing Sample Results**

	Choice-Based Estimation Sample <sup>1</sup>						Pearson
	15:15	15:30	15:60	15:75	15:120	15:150	Correlation
	(0.5)	(0.333)	(0.2)	(0.167)	(0.11)	(0.09)	Coefficients <sup>2</sup>
<b>Panel A- Unweighted Results</b>							
Training Data <sup>3</sup>							
Bankrupt (Type I Error)	0	6.67	6.67	13.33	13.33	26.67	-0.829
Nonbankrupt (Type II Error)	13.33	0	5	1.33	1.67	1.33	0.747
Overall	6.65	2.23	5.33	3.33	2.97	3.63	0.551
<b>Panel B- Weighted Results</b>							
Training Data							
Bankrupt (Type I Error)	13.33	13.33	13.33	13.33	13.33	26.67	-0.448
Nonbankrupt (Type II Error)	6.67	0	3.33	1.33	0.83	1.33	0.677
Overall	10	4.44	5.33	3.33	2.22	3.63	0.891

<sup>1</sup> Number of bankrupt : number of nonbankrupt firms in the choice-based estimation sample.

The bankrupt firms sample frequency rate(number of bankrupt firms/total number of sample firms)

<sup>2</sup> Pearson correlation coefficients between the estimation sample frequency rate and result reported in the corresponding row

<sup>3</sup> Percentage of firms misclassified

## 10.5 The Sensitivity of Optimal Cutoff Points to Misclassification Costs

The objective of this experiment is to study the sensitivity of optimal cutoff points to the misclassification costs of Type I and Type II errors in the bankruptcy prediction context. It has been shown that misclassification costs are an important factor and should be considered when the optimal cutoff points for predicting business failure models are determined. However, the costs of Type I and Type II errors are generally intangible and unmeasurable. In this experiment the Type I error cost ( $C_I$ ) and Type II error cost ( $C_{II}$ ) are not measured directly. Instead they are computed under the various ratios of  $C_I$  to  $C_{II}$ . According to Koh's suggestion [1992], the range of this ratio used to investigate the sensitivity analysis in this study is set from 1:1 to 500:1.

Four hypotheses are proposed for this experiment as follows

- $H_{26}$  : The optimal cutoff points that minimise the expected total error costs are insensitive to different misclassification costs of Type I and Type II errors for MDA method.
- $H_{27}$  : The optimal cutoff points that minimise the expected total error costs are insensitive to different misclassification costs of Type I and Type II errors for Logit method.
- $H_{28}$  : The optimal cutoff points that minimise the expected total error costs are insensitive to different misclassification costs of Type I and Type II errors for GDR method.
- $H_{29}$  : The optimal cutoff points that minimise the expected total error costs are insensitive to different misclassification costs of Type I and Type II errors for Proj method.

### 10.5.1 The Optimal Cutoff Points and Classification Accuracy for the MDA Model

The optimal cutoff points for the MDA model were calculated as follows

$$\begin{aligned} & (I^* - \mu_2)^2 / 2\sigma_2^2 - (I^* - \mu_1)^2 / 2\sigma_1^2 \\ & = \ln(\sigma_1 / \sigma_2) + \ln[\alpha_s / (1 - \alpha_s)] + \ln[(1 - \alpha_p) / \alpha_p] + \ln(C_{II} / C_I) \quad (10.5.1) \end{aligned}$$

where

- $\mu_1$  = the mean of Z scores for bankrupt firms
- $\mu_2$  = the mean of Z scores for nonbankrupt firms
- $\sigma_1$  = the standard deviation of Z scores for bankrupt firms
- $\sigma_2$  = the standard deviation of Z scores for nonbankrupt firms
- $\alpha_p$  = the proportion of bankrupt firms in the population
- $\alpha_s$  = the proportion of bankrupt firms in the sample

Under the assumption of  $\alpha_p = \alpha_s$ , Table 10.5.1 summaries the optimal cutoff points, and the number of Type I errors (NI) as well as the number of Type II errors (NII) obtained corresponding to values of  $C_I$  to  $C_{II}$  from 1:1 to 500:1. Before we discuss the results of Table 10.5.1, let us first recall the descriptive statistics on the distribution of 264 companies' Z scores reported in Appendix 3, which was computed from the MDA discriminating function. The results are repeated in Table 10.5.2 and Figure 10.5.1.

Table 10.5.2 The Statistics of Z Scores Distribution  
for MDA Method Using 264 Companies

Group	N Obs	Minimum	Maximum	Mean	Std Dev
Nonbankrupt	176	-11.22251	4.1124162	-3.62431	2.11939
Bankrupt	88	-2.4320484	17.66681	3.6237575	3.57657

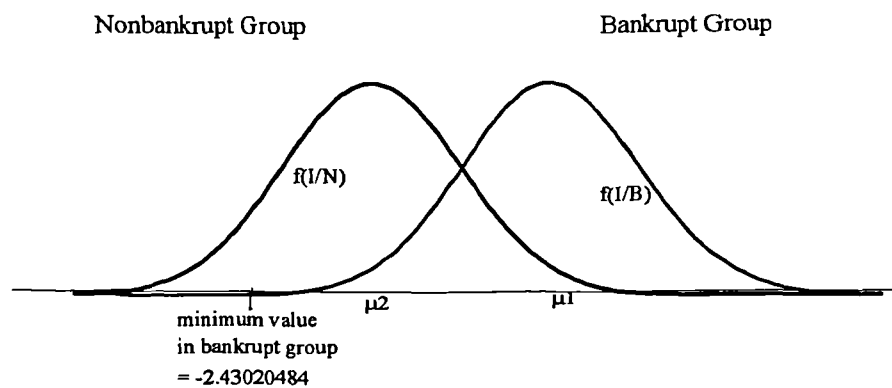


Figure 10.5.1 The Hypothetical Distribution of Z scores in MDA

As can be seen in Table 10.5.1, the numerical solution of optimal cutoff points computed from equation (10.5.1) is impossible (complex root) when the  $C_I$  to  $C_{II}$  is beyond 39:1. However, it does not matter since the solution from ratio  $C_I:C_{II} = 7:1$  is -2.6356271, which is already less than the minimum value -2.4320484 in the bankrupt group. That is, when the misclassification cost of a Type I error is more than 7 times that of a Type II



error, all observations actually in the bankrupt group should be classified into bankruptcy such that no Type I error occurs. In addition, when the misclassification ratios are 4:1, 5:1 and 6:1, the number of Type I errors is identical (1), but the number of Type II errors increases (31, 37, 47 respectively). Since there is no increase in Type I errors, there is no need to change the optimal point so as to increase the Type II errors, and thus increase the total error cost. Put another way, for 4:1, 5:1, and 6:1 cost ratios, the optimal cutoff point should remain -1.9016544 and the number of Type I errors and Type II errors will be 1 and 31 respectively in order to minimise the total misclassification cost. This outcome results from the assumption that the  $f(I/N)$  and  $f(I/B)$  meet the particular continuous distributions (normal distributions in this case) from which the numerical cutoff points were computed. It leads to an area optimal solution instead of a one point optimal solution when applying to discrete values of  $f(I/N)$  and  $f(I/B)$ . For ease of understanding, these adjusted optimal points are also called the optimal cutoff points.

The adjusted optimal points and their corresponding accuracy are presented from column 5 to 8 in Table 10.5.1. As is indicated, optimal cutoff points are affected by the misclassification costs of Type I and Type II errors. For example, the cutoff point that is optimal when  $C_I:C_{II}$  is 1:1 is not optimal when  $C_I:C_{II}$  is 10:1. However, the optimal cutoff points that minimise the expected total error costs of using the model are rather insensitive to different relative misclassification costs for the MDA method, especially when the two misclassification costs ratio is beyond 7:1.

### 10.5.2 The Optimal Cutoff Points and Classification Accuracy for the Logit Model

The optimal cutoff points for the Logit model were calculated using the following equation

$$\frac{\frac{1}{b_1} \exp(-(\frac{I^*-a_1}{b_1}))}{(1+\exp(-(\frac{I^*-a_1}{b_1}))^2)} \div \frac{\frac{1}{b_2} \exp(-(\frac{I^*-a_2}{b_2}))}{(1+\exp(-(\frac{I^*-a_2}{b_2}))^2)} = K \quad (10.5.2)$$

$$\begin{aligned} \text{where } a_1 &= 7.646775 & b_1 &= 4.54441 \\ a_2 &= -6.251775 & b_2 &= 2.2003 \end{aligned}$$

(The above values are based on the intermediate  $Z$  values which are used in the calculation of conditional probabilities  $p$ , i.e.,  $Z = \beta'X$ ,  $p=1/(1+\exp(-Z))$ , where  $\beta'$  is the vector of estimated coefficients of predictor variables,  $X$  is the vector of predictors)

For  $K$  value from 1 to 500, Table 10.5.3 summaries the numerical optimal cutoff points (on the basis of conditional probabilities  $p$ ), adjusted optimal cutoff points, and the number of Type I errors (NI) as well as the number of Type II errors (NII) corresponding to values of  $C_I$  to  $C_{II}$  from 1:1 to 500:1.

As the results indicate, the optimal cutoff points of Logit model is still not very sensitive to different levels of misclassification costs. We obtained eight different ranges of optimal decision area. They are (1) 1:1, (2) 2:1, (3) 3:1 to 4:1, (4) 5:1 to 12:1, (5) 13:1, (6) 14:1 to 29:1, (7) 30:1 to 33:1, (8) over 34:1. In the middle range the optimal cutoff point is relatively robust compared to the change ranging within 5:1. They remain the same within the relevant areas. For example, 0.08952 is the optimal cutoff point for a wide range of  $C_I:C_{II}$  ranging from 5:1 to 12:1, and 0.02710 is the optimal point for all cases where the misclassification cost of a Type I error is 14 to 29 times that of a Type II error.

Further, since NI is zero when  $C_I$  to  $C_{II}$  is 34:1, it means that when the cost of a Type I error is more than 34 times that of a Type II error, all bankrupt companies should be correctly classified because we cannot afford the consequences of this incorrect classification. We call this ratio which achieves the zero number of Type I error the critical ratio. The critical ratio 34:1 in the Logit method is much larger than that (7:1) obtained in the MDA model. In other words, the optimal cutoff points are not affected in MDA as long as the  $C_I:C_{II}$  is beyond 7:1 in contrast to the ratio 34:1 in Logit. From this viewpoint it seems that the optimal cutoff points of MDA is even more robust comparing to the Logit procedure for a wide range of relative misclassification costs.

### 10.5.3 The Optimal Cutoff Points and Classification Accuracy for the GDR Model

The optimal cutoff points for GDR were also calculated using equation (10.5.2). The parameters used are  $a_1=3.4168651$ ,  $b_1=0.980876$ ,  $a_2=-4.3621649$ ,  $b_2=0.878608$  (on the basis of the intermediate  $Z$  value).

As is shown in Table 10.5.4, the optimal cutoff point is 0.34661 for  $C_I$  to  $C_{II} = 1:1$ ; 0.27579 for  $C_I:C_{II} = 2:1$ ; 0.23868 for  $C_I$  to  $C_{II} = 3:1$ ; 0.21443 for  $C_I:C_{II}$  ranging from 4:1 to 61:1; 0.06459 for  $C_I:C_{II}$  ranging from 62:1 to 128:1; 0.04392 for  $C_I$  to  $C_{II}$  for  $C_I$  to  $C_{II}$  ranging from 129:1 to 500:1. The optimal cutoff points are relatively sensitive when the ratio of Type I error cost to that of a Type II error is within the range of 1:1 to 3:1. Nevertheless, it is quite robust to the misclassification costs ratio ranging beyond 4:1 cases. The results suggest that if the misclassification cost of Type I error is estimated to be more than 4 times that of a Type II error, the decision making using the GDR method becomes less influenced by the unequal misclassification costs due to the small change of optimal cutoff points. Meanwhile, the zero number of NI occurs at the ratio 129:1. The critical ratio in GDR is larger than that in Logit and MDA.

#### 10.5.4 The Optimal Cutoff Points and Classification Accuracy for the Proj Model

Like Logit and GDR, the optimal cutoff points for the Projection method are calculated using equation (10.5.2). The parameters used are  $a_1=0.8202149$ ,  $b_1=0.215042$ ,  $a_2=0.0242786$ ,  $b_2=0.065239$  which were developed in Chapter Nine.

From Table 10.5.5 we can see that the optimal cutoff point is 0.24018 for  $C_I$  to  $C_{II}$  ranging from 1:1 to 5:1; 0.11272 for  $C_I:C_{II}$  ranging from 6:1 to 57:1, 0.03549 for  $C_I$  to  $C_{II}$  ranging from 58:1 to 105:1, 0.02466 for  $C_I:C_{II}$  ranging from 106:1 to 362:1, 0.00967 for  $C_I:C_{II}$  ranging from 363:1 to 500:1.

By investigating the results for the Projection method, it can be concluded that the optimal cutoff points are broadly robust to different relative misclassification costs in the bankruptcy prediction context. Consequently, bankruptcy prediction models are generally applicable over a wide range of possible misclassification costs in the Projection algorithm. Further, the ratio where the number of NI is zero occurs at 363:1. This critical ratio is much larger than the ratios in MDA, Logit and GDR. However, the difference of overall accuracy between 1:1 case (95.83%) and the critical ratio case (64%) is bigger than those in the other three methods. This is shown in Table 10.5.6.

Table 10.5.1 Summary of Optimal Cutoff Points and Accuracies to Different Error Cost Ratios for MDA Method

(1) Ratio	(2) Numerical optimal cutoff	(3) NI	(4) NII	(5) Adjusted optimal cutoff	(6) Adjusted NI	(7) Adjusted NII	(8) Overall Accuracy
1:1	-0.40108	5	9	-0.40108	5	9	94.70
2:1	-1.10751	4	18	-1.10751	4	18	91.67
3:1	-1.55925	2	26	-1.55925	2	26	89.39
4:1	-1.90165	1	31	-1.90165	1	31	87.88
5:1	-2.18227	1	37	-1.90165	1	31	87.88
6:1	-2.42295	1	47	-1.90165	1	31	87.88
7:1**	-2.63563	0	53	-2.63563	0	53	79.92
8:1	-2.8276	0	61	-2.63563	0	53	79.92
9:1	-3.00366	*	*	-2.63563	0	53	79.92
10:1	-3.16715	*	*	-2.63563	0	53	79.92
11:1	-3.32048	*	*	-2.63563	0	53	79.92
12:1	-3.46549	*	*	-2.63563	0	53	79.92
13:1	-3.6036	*	*	-2.63563	0	53	79.92
14:1	-3.73592	*	*	-2.63563	0	53	79.92
15:1	-3.86338	*	*	-2.63563	0	53	79.92
16:1	-3.98674	*	*	-2.63563	0	53	79.92
17:1	-4.10665	*	*	-2.63563	0	53	79.92
18:1	-4.22366	*	*	-2.63563	0	53	79.92
19:1	-4.33827	*	*	-2.63563	0	53	79.92
20:1	-4.450912	*	*	-2.63563	0	53	79.92
21:1	-4.562	*	*	-2.63563	0	53	79.92
22:1	-4.679215	*	*	-2.63563	0	53	79.92
23:1	-4.78103	*	*	-2.63563	0	53	79.92
24:1	-4.88969	*	*	-2.63563	0	53	79.92
25:1	-4.99826	*	*	-2.63563	0	53	79.92
26:1	-5.10712	*	*	-2.63563	0	53	79.92
27:1	-5.21667	*	*	-2.63563	0	53	79.92
28:1	-5.32735	*	*	-2.63563	0	53	79.92
29:1	-5.43964	*	*	-2.63563	0	53	79.92
30:1	-5.55413	*	*	-2.63563	0	53	79.92
31:1	-5.6715	*	*	-2.63563	0	53	79.92
32:1	-5.79261	*	*	-2.63563	0	53	79.92
33:1	-5.91856	*	*	-2.63563	0	53	79.92
34:1	-6.05086	*	*	-2.63563	0	53	79.92
35:1	-6.19163	*	*	-2.63563	0	53	79.92
36:1	-6.34411	*	*	-2.63563	0	53	79.92
37:1	-6.51379	*	*	-2.63563	0	53	79.92
38:1	-6.7114	*	*	-2.63563	0	53	79.92
39:1	-6.9652	*	*	-2.63563	0	53	79.92
40:1	complex root	*	*	-2.63563	0	53	79.92
41:1	complex root	*	*	-2.63563	0	53	79.92
:	:	:	:	:	:	:	:
500:1	complex root	*	*	-2.63563	0	53	79.92

\*\* denotes the critical ratio

Table 10.5.3 Summary of Optimal Cutoff Points and Accuraies to Different Error Cost Ratios  
for Logit Method

Ratio	Numerical optimal cutoff	Adjusted optimal cutoff	Adjusted NI	Adjusted NII	Overall accuracy	Ratio	Numerical optimal cutoff	Adjusted optimal cutoff	Adjusted NI	Adjusted NII	Overall accuracy	Ratio	Numerical optimal cutoff	Adjusted optimal cutoff	Adjusted NI	Adjusted NII	Overall accuracy
1:1	0.37271	0.37271	10	11	92.0	21:1	0.02	0.02710	2	53	79.2	41:1	0.00553	0.00761	0	77	70.8
2:1	0.21700	0.21700	8	14	91.7	22:1	0.01	0.02710	2	53	79.2	42:1	0.00529	0.00761	0	77	70.8
3:1	0.14982	0.14982	6	16	91.7	23:1	0.01	0.02710	2	53	79.2	43:1	0.00507	0.00761	0	77	70.8
4:1	0.11281	0.14982	6	16	91.7	24:1	0.01	0.02710	2	53	79.2	44:1	0.00486	0.00761	0	77	70.8
5:1	0.08952	0.08952	5	28	87.5	25:1	0.01	0.02710	2	53	79.2	45:1	0.00466	0.00761	0	77	70.8
6:1	0.07359	0.08952	5	28	87.5	26:1	0.01	0.02710	2	53	79.2	46:1	0.00447	0.00761	0	77	70.8
7:1	0.06204	0.08952	5	28	87.5	27:1	0.01	0.02710	2	53	79.2	47:1	0.00429	0.00761	0	77	70.8
8:1	0.05332	0.08952	5	28	87.5	28:1	0.01	0.02710	2	53	79.2	48:1	0.00412	0.00761	0	77	70.8
9:1	0.04651	0.08952	5	28	87.5	29:1	0.01	0.02710	2	53	79.2	49:1	0.00396	0.00761	0	77	70.8
10:1	0.04105	0.08952	5	28	87.5	30:1	0.00930	0.00930	1	72	72.3	50:1	0.00381	0.00761	0	77	70.8
11:1	0.03660	0.08952	5	28	87.5	31:1	0.01	0.00930	1	72	72.3	51:1	0.00366	0.00761	0	77	70.8
12:1	0.03289	0.08952	5	28	87.5	32:1	0.00840	0.00930	1	72	72.3	52:1	0.00352	0.00761	0	77	70.8
13:1	0.02977	0.02977	4	47	80.7	33:1	0.01	0.00930	1	72	72.3	53:1	0.00338	0.00761	0	77	70.8
14:1	0.02710	0.02710	2	53	79.2	34:1**	0.01	0.01	0	77	70.8	54:1	0.00325	0.00761	0	77	70.8
15:1	0.02480	0.02710	2	53	79.2	35:1	0.01	0.01	0	77	70.8	55:1	0.00313	0.00761	0	77	70.8
16:1	0.02280	0.02710	2	53	79.2	36:1	0.01	0.01	0	77	70.8	56:1	0.00301	0.00761	0	77	70.8
17:1	0.02104	0.02710	2	53	79.2	37:1	0.01	0.01	0	77	70.8	57:1	0.00290	0.00761	0	77	70.8
18:1	0.01949	0.02710	2	53	79.2	38:1	0.01	0.01	0	77	70.8	58:1	0.00000	0.00761	0	77	70.8
19:1	0.01811	0.02710	2	53	79.2	39:1	0.01	0.01	0	77	70.8	:	:	:	:	:	:
20:1	0.01688	0.02710	2	53	79.2	40:1	0.01	0.01	0	77	70.8	500:1	0.00000	0.00761	0	77	70.8

\*\* denotes the critical ratio

Table 10.5.4 Summary of Optimal Cutoff Points and Accuracies to Different Error Cost Ratios  
for GDR Method

Ratio	Numerical optimal cutoff	Adjusted optimal cutoff	Adjusted NI	Adjusted NII	Overall accuracy	Ratio	Numerical optimal cutoff	Adjusted optimal cutoff	Adjusted NI	Adjusted NII	Overall accuracy	Ratio	Numerical optimal cutoff	Adjusted optimal cutoff	Adjusted NI	Adjusted NII	Overall accuracy
1:1	0.34661	0.34661	6	8	94.70	21:1	0.10791	0.21443	2	16	93.18	63:1	0.06407	0.06459	1	45	80.58
2:1	0.27579	0.27579	4	13	93.56	22:1	0.10568	0.21443	2	16	93.18	64:1	0.06357	0.06459	1	45	80.58
3:1	0.23868	0.23868	3	14	93.56	23:1	0.10358	0.21443	2	16	93.18	65:1	0.06307	0.06459	1	45	80.58
4:1	0.21443	0.21443	2	16	93.18	24:1	0.10160	0.21443	2	16	93.18	66:1	0.06259	0.06459	1	45	80.58
5:1	0.19683	0.21443	2	16	93.18	25:1	0.09972	0.21443	2	16	93.18	67:1	0.06211	0.06459	1	45	80.58
6:1	0.18324	0.21443	2	16	93.18	26:1	0.09795	0.21443	2	16	93.18	68:1	0.06165	0.06459	1	45	80.58
7:1	0.17229	0.21443	2	16	93.18	27:1	0.09627	0.21443	2	16	93.18	69:1	0.06119	0.06459	1	45	80.58
8:1	0.16320	0.21443	2	16	93.18	28:1	0.09467	0.21443	2	16	93.18	:	:	:	:	:	:
9:1	0.15549	0.21443	2	16	93.18	29:1	0.09314	0.21443	2	16	93.18	128:1	0.04411	0.06459	1	45	80.58
10:1	0.14883	0.21443	2	16	93.18	30:1	0.09169	0.21443	2	16	93.18	129:1**	0.04392	0.04392	0	54	79.55
11:1	0.14299	0.21443	2	16	93.18	:	:	:	:	:	:	130:1	0.04373	0.04392	0	54	79.55
12:1	0.13782	0.21443	2	16	93.18	61:1	0.06512	0.21443	2	16	93.18	131:1	0.04355	0.04392	0	54	79.55
13:1	0.13318	0.21443	2	16	93.18	62:1	0.06459	0.06459	1	45	80.58	132:1	0.04337	0.04392	0	54	79.55
14:1	0.12901	0.21443	2	16	93.18	63:1	0.06407	0.06459	1	45	80.58	133:1	0.04318	0.04392	0	54	79.55
15:1	0.12521	0.21443	2	16	93.18	64:1	0.06357	0.06459	1	45	80.58	134:1	0.04300	0.04392	0	54	79.55
16:1	0.12173	0.21443	2	16	93.18	65:1	0.06307	0.06459	1	45	80.58	135:1	0.04283	0.04392	0	54	79.55
17:1	0.11854	0.21443	2	16	93.18	66:1	0.06259	0.06459	1	45	80.58	136:1	0.04265	0.04392	0	54	79.55
18:1	0.11559	0.21443	2	16	93.18	67:1	0.06211	0.06459	1	45	80.58	137:1	0.04248	0.04392	0	54	79.55
19:1	0.11285	0.21443	2	16	93.18	68:1	0.06165	0.06459	1	45	80.58	:	:	:	:	:	:
20:1	0.11030	0.21443	2	16	93.18	69:1	0.06119	0.06459	1	45	80.58	500:1	0.01841	0.04392	0	54	79.55

\*\* denotes the critical ratio

Table 10.5.5 Summary of Optimal Cutoff Points and Accuracies to Different Error Cost Ratios  
for Proj Method

Ratio	Numerical optimal cutoff	Adjusted optimal cutoff	Adjusted NI	Adjusted NII	Overall accuracy	Ratio	Numerical optimal cutoff	Adjusted optimal cutoff	Adjusted NI	Adjusted NII	Overall accuracy	Ratio	Numerical optimal cutoff	Adjusted optimal cutoff	Adjusted NI	Adjusted NII	Overall accuracy
1:1	0.24018	0.24018	5	6	95.83	21:1	0.06147	0.11272	3	19	91.67	106:1	0.02466	0.02466	1	65	75
2:1	0.18212	0.24018	5	6	95.83	22:1	0.06002	0.11272	3	19	91.67	107:1	0.02451	0.02466	1	65	75
3:1	0.15343	0.24018	5	6	95.83	23:1	0.05866	0.11272	3	19	91.67	108:1	0.02437	0.02466	1	65	75
4:1	0.13531	0.24018	5	6	95.83	24:1	0.05738	0.11272	3	19	91.67	109:1	0.02422	0.02466	1	65	75
5:1	0.12246	0.24018	5	6	95.83	25:1	0.05618	0.11272	3	19	91.67	110:1	0.02408	0.02466	1	65	75
6:1	0.11272	0.11272	3	19	91.67	26:1	0.05504	0.11272	3	19	91.67	111:1	0.02394	0.02466	1	65	75
7:1	0.10498	0.11272	3	19	91.67	27:1	0.05396	0.11272	3	19	91.67	112:1	0.02380	0.02466	1	65	75
8:1	0.09863	0.11272	3	19	91.67	28:1	0.05294	0.11272	3	19	91.67	113:1	0.02367	0.02466	1	65	75
9:1	0.09330	0.11272	3	19	91.67	:	:	:	:	:	:	114:1	0.02353	0.02466	1	65	75
10:1	0.08873	0.11272	3	19	91.67	57:1	0.03584	0.11272	3	19	91.67	115:1	0.02340	0.02466	1	65	75
11:1	0.08475	0.11272	3	19	91.67	58:1	0.03549	0.03549	2	46	81.82	116:1	0.02327	0.02466	1	65	75
12:1	0.08125	0.11272	3	19	91.67	59:1	0.03514	0.03549	2	46	81.82	117:1	0.02314	0.02466	1	65	75
13:1	0.07814	0.11272	3	19	91.67	60:1	0.03480	0.03549	2	46	81.82	118:1	0.02301	0.02466	1	65	75
14:1	0.07535	0.11272	3	19	91.67	61:1	0.03447	0.03549	2	46	81.82	119:1	0.02289	0.02466	1	65	75
15:1	0.07282	0.11272	3	19	91.67	62:1	0.03414	0.03549	2	46	81.82	:	:	:	:	:	:
16:1	0.07052	0.11272	3	19	91.67	63:1	0.03383	0.03549	2	46	81.82	362:1	0.00970	0.02466	1	65	75
17:1	0.06841	0.11272	3	19	91.67	64:1	0.03352	0.03549	2	46	81.82	363:1**	0.00967	0.00967	0	95	64
18:1	0.06647	0.11272	3	19	91.67	65:1	0.03322	0.03549	2	46	81.82	:	:	:	:	:	:
19:1	0.06468	0.11272	3	19	91.67	:	:	:	:	:	:	:	:	:	:	:	:
20:1	0.06302	0.11272	3	19	91.67	105:1	0.02481	0.03549	2	46	81.82	500:1	0.00688	0.00967	0	95	64

\*\* denotes the critical ratio

**Table 10.5.6 The Difference of Overall Accuracy  
between 1:1 Ratio and Critical Ratio for four Methods**

Method	Critical ratio	Overall accuracy	Overall accuracy	Difference
MDA	7:1	94.70%	79.92%	14.78%
Logit	34:1	92.00%	70.80%	21.20%
GDR	129:1	94.70%	79.55%	15.15%
Proj	363:1	95.83%	64.00%	31.83%

### **10.5.5 The Comparison of Type I Error to Different Relative Misclassification Costs among the Four Methods**

When making a comparison of the sensitivity of optimal cutoff points to different relative error costs among four methods through graphics analysis, we should put the alternative ratios of Type I error to Type II error on an axis of abscissas (X axis), and the corresponding optimal cutoff points obtained in each method on an axis of ordinates (Y axis). However, because the optimal cutoff points do not have the same scaling between ANNs & Logit (always between 0 and 1) and MDA (they depend on the distribution of Z scores, and not necessarily between 0 and 1), comparing the degree of robustness among these four techniques on one figure is impossible. Fortunately, the concept of the number of Type I errors generated by corresponding optimal cutoff points is similar to the optimal cutoff point itself and has the same scaling for four methods. Hence, we used the number of Type I errors as the measure on the Y axis instead of optimal cutoff points. Figure 10.5.2 displays the comparison of sensitivity for MDA, Logit, GDR and Proj.

As is indicated in Figure 10.5.2, the number of Type I errors (optimal point) in the Logit method has more different ranges (8), and thus is slightly more sensitive than those in the other three methods. But it is still rather robust to different relative misclassification costs. On the other hand, the MDA provides the fastest convergence to the zero number of Type I error. The optimal cutoff points are not affected when the misclassification cost of a Type I error is more than 7 times that of a Type II error.



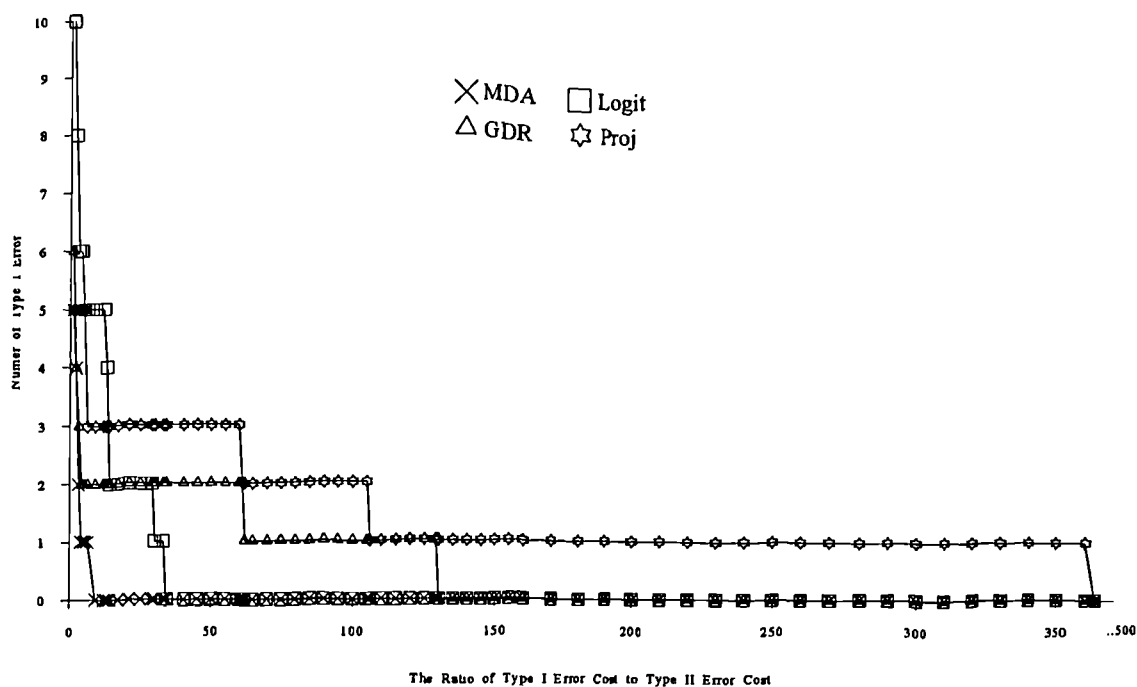


Figure 10.5.2 The Comparison of Type I Error to Different Relative Misclassification Costs among the Four Methods

Generally, for these four discriminating techniques, the optimal cutoff points are relatively sensitive within the low cost ratio area, but are applicable over a wide range of possible misclassification costs. Thus, if the Type I error cost is expected to be much higher than the Type II error cost, the slightly inaccurate estimation of two types of error costs does not appear to be a very serious limitation. Otherwise, nonconsideration of misclassification costs in the prediction model can lead to different optimal cutoff points, thus and to erroneous classification results.

## 10.6 The Influence of Different Base Rates between Training and Testing Data on Classification Accuracy among the Four Methods

The purpose of this study is to assess if the predictive accuracy of alternative discriminating methods could be affected when the base rate (the proportion of bankrupt firms to nonbankrupt firms) differs between the training sample and the testing sample.

The base rate may have an impact on a prediction technique's performance in two ways; First, a technique may not work well when the firms of interest constitute a very small percentage of the population due to an inability to identify the features necessary for classification. Second, if a classification model based on a certain base rate works well across other proportions, it is feasible to build a model using different proportions of cases of interest from those actually occurring in the present population.

The relevant hypotheses to be tested in this experiment are stated as follows:

- $H_{30}$  : There is no difference in predictive capability using different base rates between training and testing data composition in the MDA technique.
- $H_{31}$  : There is no difference in predictive capability using different base rates between training and testing data composition in the Logit technique.
- $H_{32}$  : There is no difference in predictive capability using different base rates between training and testing data composition in the GDR technique.
- $H_{33}$  : There is no difference in predictive capability using different base rates between training and testing data composition in the Projection technique.
- $H_{34}$  : The statistical methods perform as well as the neural networks when the base rates of the training sample and testing sample are different.
- $H_{35}$  : The statistical methods are more robust than the neural networks when the base rates of the training sample and testing sample are different.

### 10.6.1 The Results and Analyses

#### Training Data — "Learning"

Table 10.6.1 displays the learning performances of the four alternative methods. These outcomes are averaging the misclassification rates of 20 replications for each of nine combinations of the following two-factor design.

Sample size 120	Testing set	Testing set	Testing set
Training set base rate	60/60 60/60	60/60 20/100	60/60 12/108
Training set base rate	20/100 60/60	20/100 20/100	20/100 20/100
Training set base rate	12/108 60/60	12/108 20/100	12/108 12/108

**Table 10.6.1 Misclassification Rates of Different Base Rates  
between Training and Testing Data Compositions  
Using Training Samples**

Type I Error				
Composition	Method	1/1 in Testing	1/5 in Testing	1/9 in Testing
1/1 in Training	MDA	8.58	8.58	8.58
	Logit	6.75	6.75	6.75
	GDR	6.67	7.84	8.42
	Proj	4.58	5.00	5.08
1/5 in Training	MDA	10.75	10.75	10.75
	Logit	4.25	4.25	4.25
	GDR	19.00	21.00	16.08
	Proj	12.67	13.75	10.17
1/9 in Training	MDA	10.00	10.00	10.00
	Logit	5.83	5.83	5.83
	GDR	35.83	35.42	25.83
	Proj	17.50	18.67	12.08
Type II Error				
Composition	Method	1/1 in Testing	1/5 in Testing	1/9 in Testing
1/1 in Training	MDA	4.42	4.42	4.42
	Logit	5.91	5.91	5.91
	GDR	6.09	4.35	4.21
	Proj	3.50	3.22	3.08
1/5 in Training	MDA	2.05	2.05	2.05
	Logit	0.80	0.80	0.80
	GDR	0.92	0.60	0.45
	Proj	0.90	0.70	0.20
1/9 in Training	MDA	2.45	2.45	2.45
	Logit	0.83	0.83	0.83
	GDR	0.37	0.56	0.42
	Proj	0.51	0.46	0.28
Overall Error				
Composition	Method	1/1 in Testing	1/5 in Testing	1/9 in Testing
1/1 in Training	MDA	6.50	6.50	6.50
	Logit	6.33	6.33	6.33
	GDR	6.38	6.09	6.32
	Proj	4.04	4.11	4.08
1/5 in Training	MDA	3.50	3.50	3.50
	Logit	1.37	1.37	1.37
	GDR	3.93	4.00	3.06
	Proj	2.86	2.87	1.86
1/9 in Training	MDA	3.21	3.21	3.21
	Logit	1.33	1.33	1.33
	GDR	3.92	4.04	2.96
	Proj	2.21	2.28	1.46

\* The number in the table are indicated in percentage

Investigating the results, three interesting findings may be summarised

- (1) In terms of STMs, it is shown that the misclassification rates for training data are identical, regardless of combinations with the different testing composition data, as long as the base rate in the training samples is the same (i.e., the same row results). On the other hand, the neural networks produce different performances with the same learning data when matched with different testing samples.

This difference results from the distinct estimation processes between them. The STMs have a systematic discriminating algorithm. The estimation process in these methods assumes that a fixed set of learning data is given, and thus yields same results provided that the learning data is the same. However, the ANNs are basically unstructured methodology based on the range of whole data (i.e., both the training and testing data). Additionally, the data is assumed to come into the system sequentially or randomly, and this may provide different results even if the learning samples are identical. In this experiment, though all the parameters are set identically in each case when running a neural network task, the application results using the same training samples still vary in terms of different testing data, because the length and scale of the input vector (including training and testing data) will change from case to case. In other words, at the learning stage, the MDA and Logit procedures are not influenced by the contents of the testing sample. By contrast, the learning results for neural networks are not independent of the value of testing data even if other parameters hold constant. From the viewpoint of the information systems practitioner, the problem of similitude or replicability is one of several worrying concerns with the use of ANNs.

- (2) As we pointed out above, in ANNs, the training performance may be affected by the different testing data. However, this impact is not of the type we commonly expect. Normally, we may guess that the performance will be best if the base rate between training and testing is equal. That is, when the training base rate is fixed to 1 to 1 (i.e., 1/1), the highest classification should be obtained when combining with the 1/1 instead of 1/5 or 1/9 testing base rate. Likewise, when the training base rate is fixed to 1/5 (or 1/9), the highest classification accuracy should be obtained when combining with the 1/5 (or 1/9) base rate of testing. Observing the

results in Table 10.6.1, we are aware that this is not the case. This phenomenon is more clearly shown in Figures 10.6.1 to 10.6.9. Irrespective of the Type I, Type II or Overall error, it seems that we cannot conclude that neural networks have the best learning when the proportion of two groups between training and testing is identical.

- (3) Of the statistical methods, the Logit has better predictive ability than the MDA for almost every combination cell (except for 1/1 in training case). As for ANNs, the Projection algorithm always yields lower error rates than the GDR approach across all cases. However the superiority of statistical methods or ANNs depends on the composition of the two groups between the training and testing data.

### Testing Data — "Generalisation"

In this experiment the generalisation for each technique is probably our chief concern. Let us now examine the results for testing data. Table 10.6.2 reports the generalisation capacities of the four methods using various combinations of base rates between the training and testing samples.

First, we explore the influence of different base rates on each method. Figures 10.6.10 to 10.6.13 illustrate the results for MDA, Logit, GDR and Projection respectively. Three points are worth mentioning:

- (1) For various base rates in learning and testing, the best performance usually occurs when the base rate between learning and testing sample is equal (i.e., 1/1 vs. 1/1, 1/5 vs. 1/5 and 1/9 vs. 1/9). These situations are circled in Figures 10.6.10 to 10.6.13 for each of the four methods individually. In addition, observing further these outcomes with the identical base rates in training and testing data, we notice that the prediction power on 1/1 base rate is always worse than that on 1/5 base rate, and also worse than that on 1/9 base rate in terms of overall error rates. However, there is no significant difference between 1/5 cases and 1/9 cases over the four methods. These results indicate that not only the statistical methods, but also the ANNs, do not provide better understanding and differentiation between two groups when an equal number of examples, compared to an imbalanced proportion, of different groups in the sample is used. This finding is consistent

**Table 10.6.2 Misclassification Rates of Different Base Rates  
between Training and Testing Data Compositions  
Using Testing Samples**

Type I Error				
Composition	Method	1/1 in Testing	1/5 in Testing	1/9 in Testing
1/1 in Training	MDA	11.75	10.75	10.42
	Logit	6.33	4.75	5.83
	GDR	10.17	9.75	10.42
	Proj	5.58	5.50	7.08
1/5 in Training	MDA	17.83	18.25	20.96
	Logit	15.50	14.00	17.08
	GDR	28.83	27.00	25.83
	Proj	16.50	16.25	21.67
1/9 in Training	MDA	19.08	16.00	22.08
	Logit	18.83	14.00	22.92
	GDR	35.75	32.25	30.25
	Proj	21.00	19.00	21.67
Type II Error				
Composition	Method	1/1 in Testing	1/5 in Testing	1/9 in Testing
1/1 in Training	MDA	6.25	6.10	6.07
	Logit	6.75	7.25	6.85
	GDR	6.42	4.95	4.94
	Proj	6.33	6.53	6.57
1/5 in Training	MDA	3.58	2.90	3.10
	Logit	3.08	2.55	2.59
	GDR	1.17	1.15	1.02
	Proj	1.92	1.85	1.48
1/9 in Training	MDA	3.42	2.90	2.91
	Logit	1.92	1.65	2.22
	GDR	0.83	1.10	0.74
	Proj	1.42	1.30	1.25
Overall Error				
Composition	Method	1/1 in Testing	1/5 in Testing	1/9 in Testing
1/1 in Training	MDA	9.00	6.87	6.50
	Logit	6.54	6.83	6.75
	GDR	8.29	5.75	5.49
	Proj	5.96	6.36	6.62
1/5 in Training	MDA	10.71	5.46	4.89
	Logit	9.29	4.46	4.04
	GDR	15.00	5.46	3.50
	Proj	9.21	4.25	3.50
1/9 in Training	MDA	11.25	5.08	4.83
	Logit	10.38	3.71	4.29
	GDR	18.29	6.29	3.69
	Proj	11.21	4.25	3.29

\* The numbers in the table are indicated in percentage

with the earlier results but contradicts one's intuition. It implies that a one to one matching design is not the best choice if a matching approach is applied.

- (2) Generally, with different base rates in training and testing, the closer they are, the lower the misclassification can be achieved for both STMs and ANNs.
- (3) There is higher predictive power when using a closer base rate of training and a farther base rate of testing than when using a farther base rate of training and a closer base rate of testing. For example, if we apply the classification rule generated by 1/5 base rate of training data to the testing data with 1/9 base rate, we can expect that predictive performance is better than the reverse situation (i.e., the 1/9 base rate in training and the 1/5 base rate in testing). This is because the classification rule (cutoff point) is previously distorted by the training sets which consist of farther proportions of two groups. The distorted cutoff point then yields more inaccurate classification when applied to the testing sample with the closer proportion of two groups.

After analysing the results of misclassification for each method under various combinations of base rates between training and testing, we will now compare the three types of errors among four techniques. Figures 10.6.14 to 10.6.22 illustrate the Type I, Type II and Overall error rates for all nine cases. We summarise several important conclusions below:

- (1) In every combination four methods appear to predict nonbankrupt firms quite well.

That is, the Type II errors are generally low across all cases. However, when the proportion of failing firms is much smaller than the proportion of nonfailing firms, the statistical methods on average produced a better performance than neural networks in predicting bankruptcy. In this situation the ANNs, especially for the GDR algorithm, seem to focus more on the overall performance instead of balancing the misclassification of two groups, by adjusting the weights so that the total error function minimised is strongly affected by the discordant proportion of two categories in the sample. Therefore, to minimise the Overall error, ANNs sometimes reach this goal at the cost of magnifying the error rates of the group which has the relatively smaller proportion in the sample.

- (2) However, for a 1/1 base rate in training data, the testing outcomes in ANNs have lower misclassifications than those in the statistical methods for all Type I, Type II and Overall error rates. This result is consistent with the findings of Wilson and Sharda [1994] that with a balanced proportion of two groups, the neural networks provide better classification.
- (3) In terms of generalisation, for the statistical methods, the Logit extensively performs better than MDA. While in ANNs, the Projection algorithm always provides lower misclassification rates than the GDR approach.
- (4) If the relative misclassification costs of Type I to Type II errors are taken into account for the bankruptcy prediction model, the difference between Type I and Type II error rates becomes important. This information is reported in Table 10.6.3.

**Table 10.6.3 The Difference between Type I and Type II Errors  
Over all Combinations**

Composition	Method	Type I Error		
		1/1 in Testing	1/5 in Testing	1/9 in Testing
1/1 in Training	MDA	5.50	4.65	4.35
	Logit	-0.42	-2.50	-1.02
	GDR	3.75	4.80	5.48
	Proj	-0.75	-1.03	0.51
1/5 in Training	MDA	14.25	15.35	17.86
	Logit	12.42	11.45	14.49
	GDR	27.66	25.85	24.81
	Proj	14.58	14.58	20.19
1/9 in Training	MDA	15.66	13.10	19.17
	Logit	16.91	12.35	20.70
	GDR	34.92	31.15	29.51
	Proj	19.58	17.7	20.42

As indicated in the results, the largest difference was achieved in the GDR method, while the smallest difference almost always occurs in the Logit procedure. Generally the disparity of the two type errors in statistical methods, owing to different compositions of base rate between training and testing, are less than ANNs', although the overall performance is not necessarily better. In particular, for the Logit procedure, it is observed



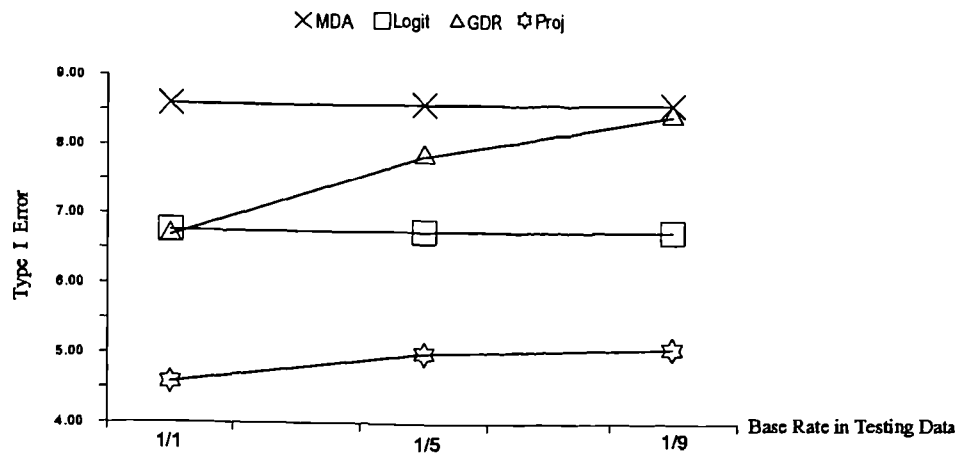


Figure 10.6.1 Type I Error vs. Variuos Base Rates of Testing Data when Base Rate of Training Data is Fixed to 1/1 Using Training Data Results

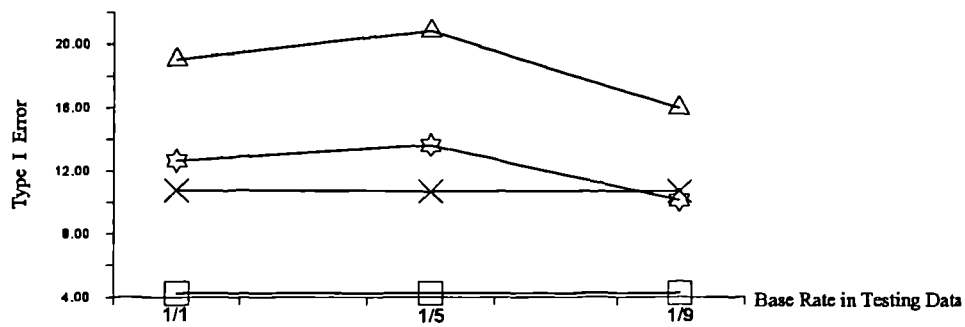


Figure 10.6.2 Type I Error vs. Various Base Rates of Testing Data when Base Rate of Training Data is Fixed to 1/5 Using Training Data Results

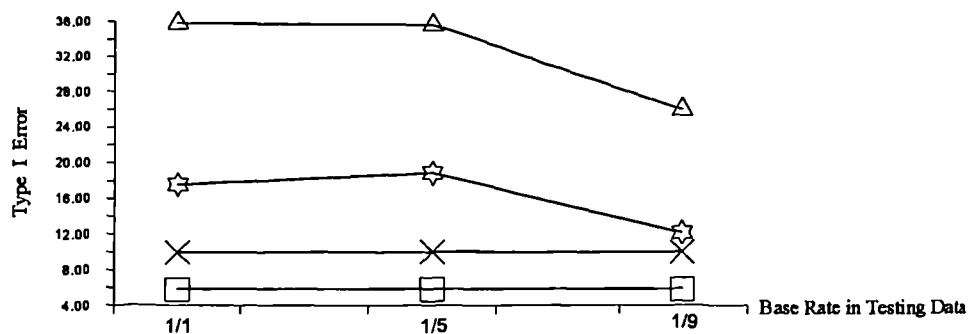


Figure 10.6.3 Type I Error vs. Various Base Rates of Testing Data when Base Rate of Training Data is Fixed to 1/9 Using Training Data Results

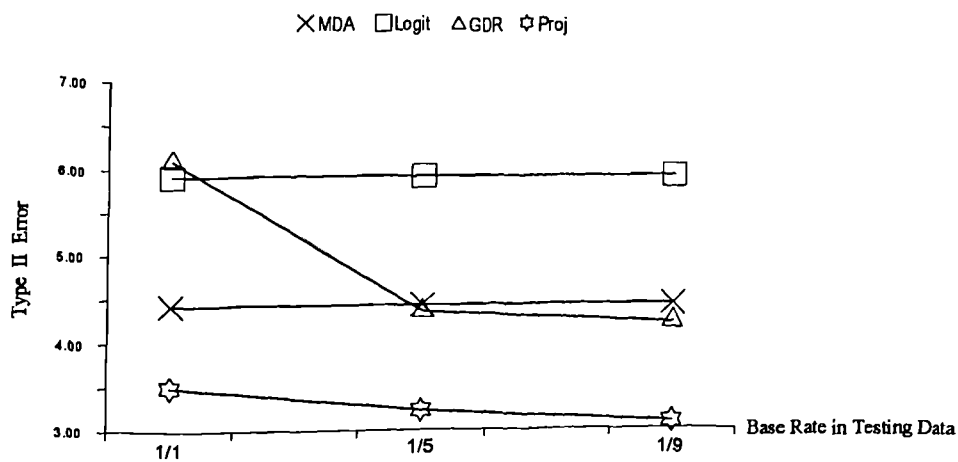


Figure 10.6.4 Type II Error vs. Various Base Rates of Testing Data when Base Rate of Training Data is Fixed to 1/1 Using Training Data Results

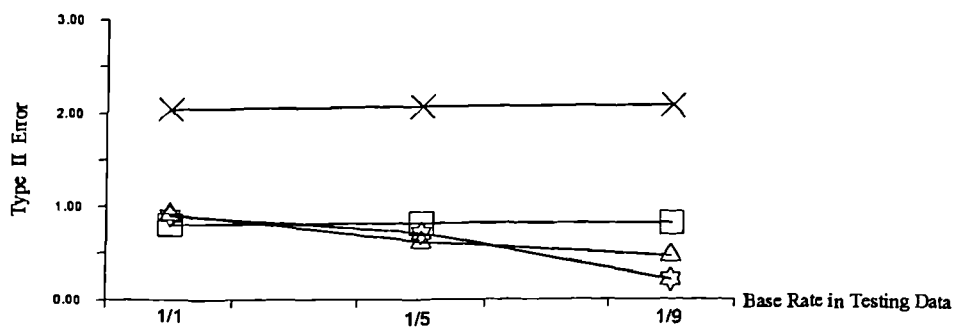


Figure 10.6.5 Type II Error vs. Various Base Rates of Testing Data when Base Rate of Training Data is Fixed to 1/5 Using Training Data Results

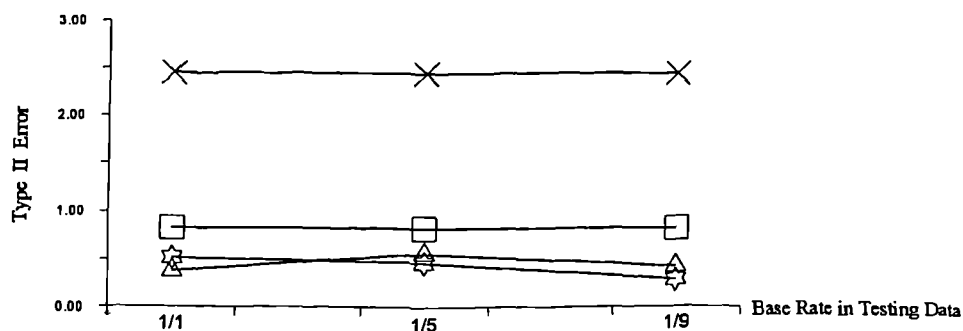


Figure 10.6.6 Type II Error vs. Various Base Rates of Testing Data when Base Rate of Training Data is Fixed to 1/9 Using Training Data Results

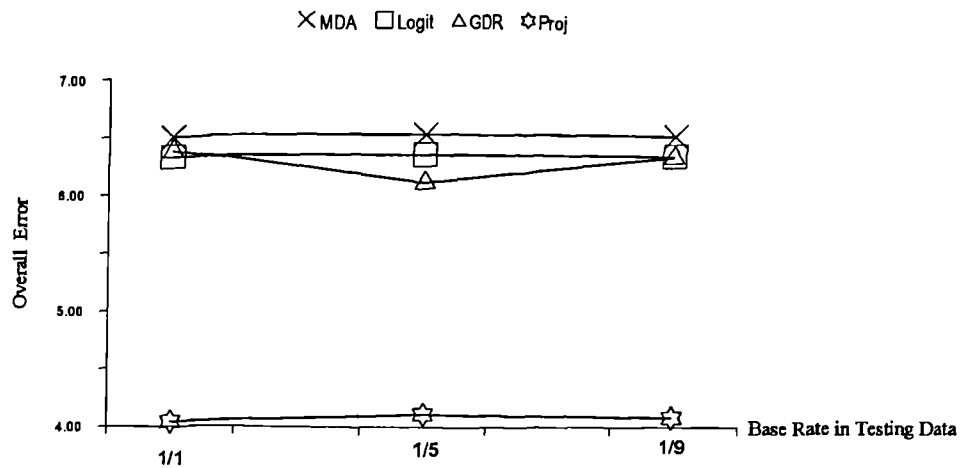


Figure 10.6.7 Overall Error vs. Variuos Base Rates of Testing Data  
when Base Rate of Training Data is Fixed to 1/1 Using Training Data Results

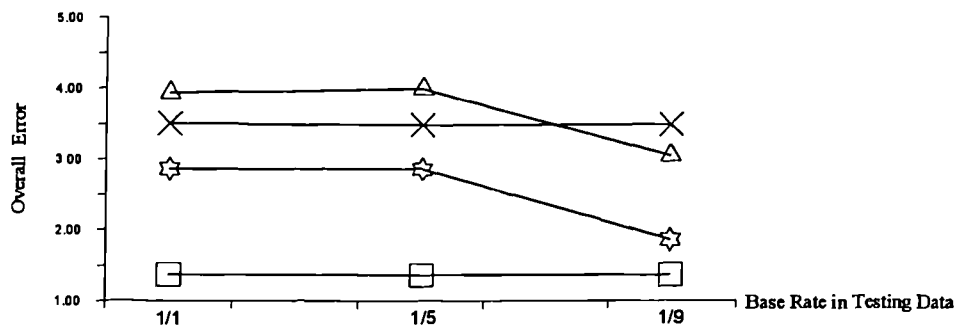


Figure 10.6.8 Overall Error vs. Various Base Rates of Testing Data  
when Base Rate of Training Data is Fixed to 1/5 Using Training Data Results

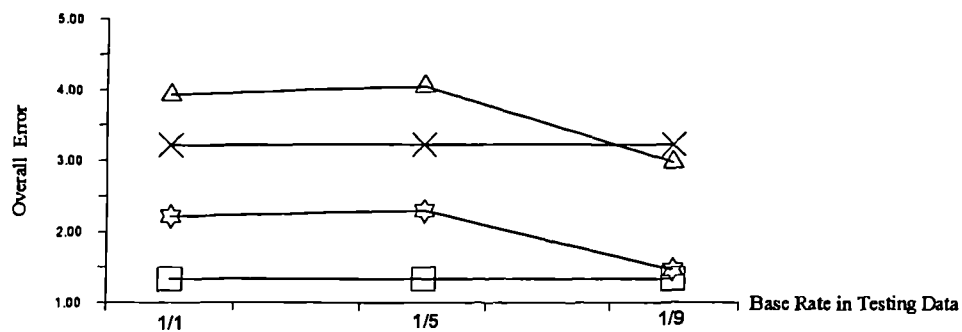


Figure 10.6.9 Overall Error vs. Various Base Rates of Testing Data  
when Base Rate of Training Data is Fixed to 1/9 Using Training Data Results

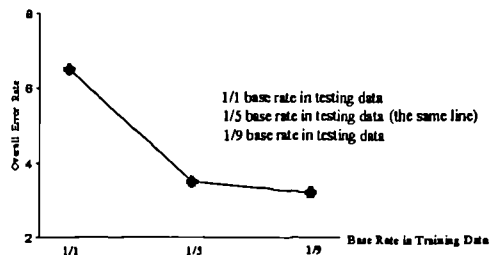


Figure 10.6.I Overall Error for Various Base Rates between Training and Testing Data in MDA Using Training Data Results

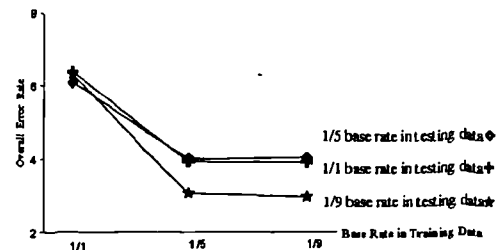


Figure 10.6.III Overall Error for Various Base Rates between Training and Testing Data in GDR Using Training Data Results

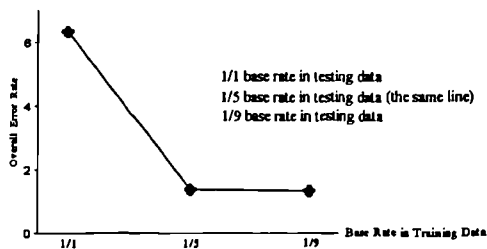


Figure 10.6.II Overall Error for Various Base Rates between Training and Testing Data in Logit Using Training Data Results

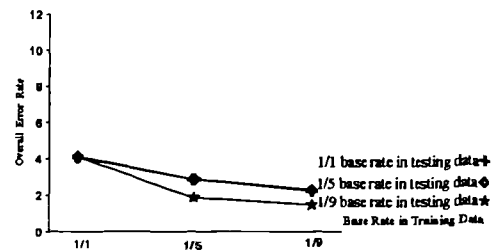


Figure 10.6.IV Overall Error for Various Base Rates between Training and Testing Data in Proj Using Training Data Results

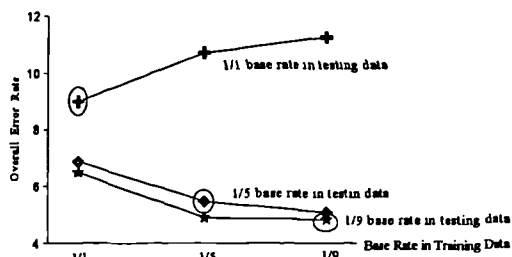


Figure 10.6.10 Overall Error for Various Base Rates between Training and Testing Data in MDA Using Testing Data Results

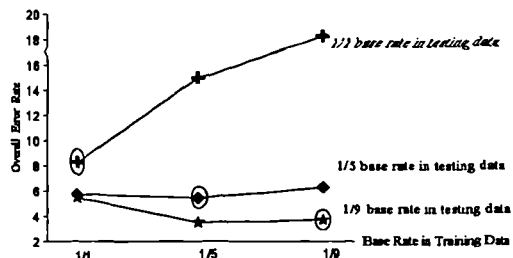


Figure 10.6.12 Overall Error for Various Base Rates between Training and Testing Data in GDR Using Testing Data Results

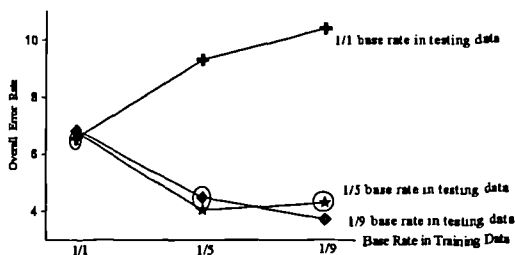


Figure 10.6.11 Overall Error for Various Base Rates between Training and Testing Data in Logit Using Testing Data Results

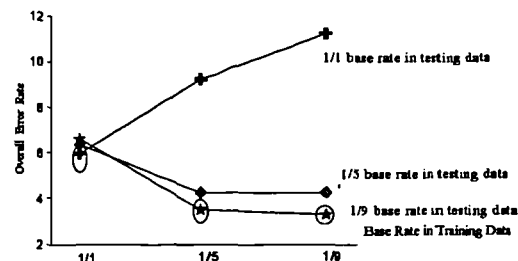


Figure 10.6.13 Overall Error for Various Base Rates between Training and Testing Data in Proj Using Testing Data Results

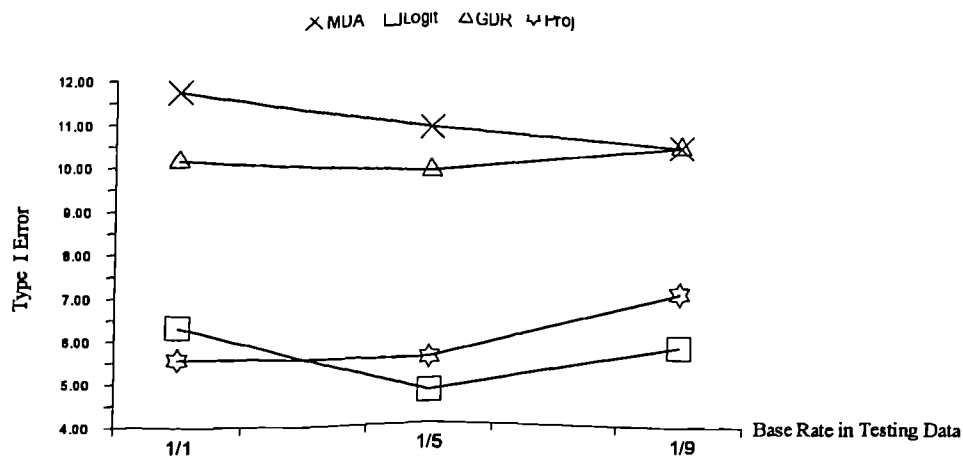


Figure 10.6.14 Type I Error vs. Various Base Rates of Testing Data when Base Rate of Training Data is Fixed to 1/1 Using Testing Data Results

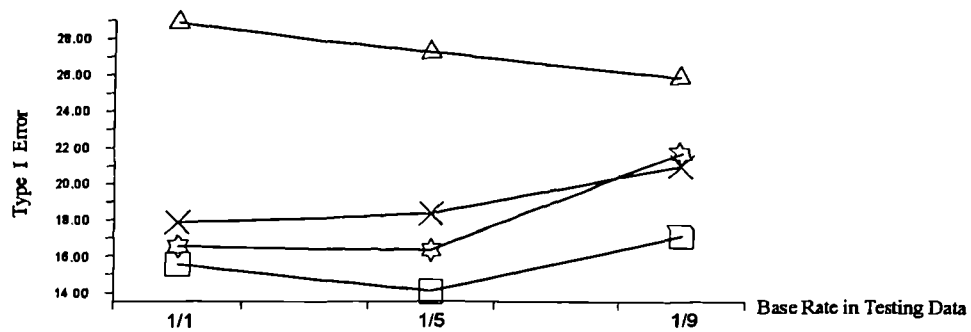


Figure 10.6.15 Type I Error vs. Various Base Rates of Testing Data when Base Rate of Training Data is Fixed to 1/5 Using Testing Data Results

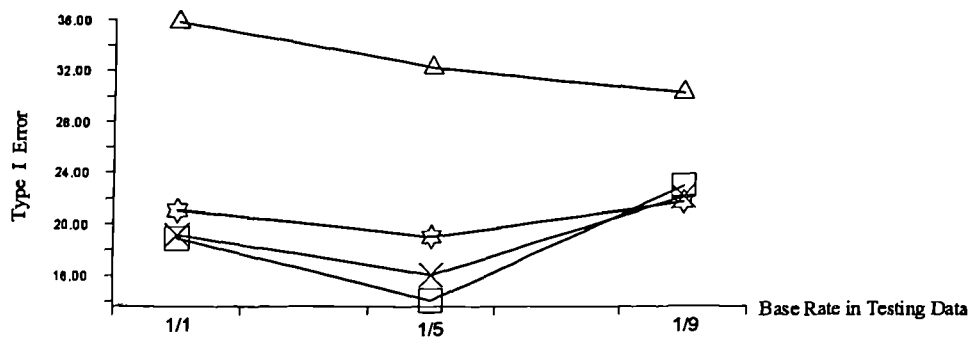


Figure 10.6.16 Type I Error vs. Various Base Rates of Testing Data when Base Rate of Training Data is Fixed to 1/9 Using Testing Data Results

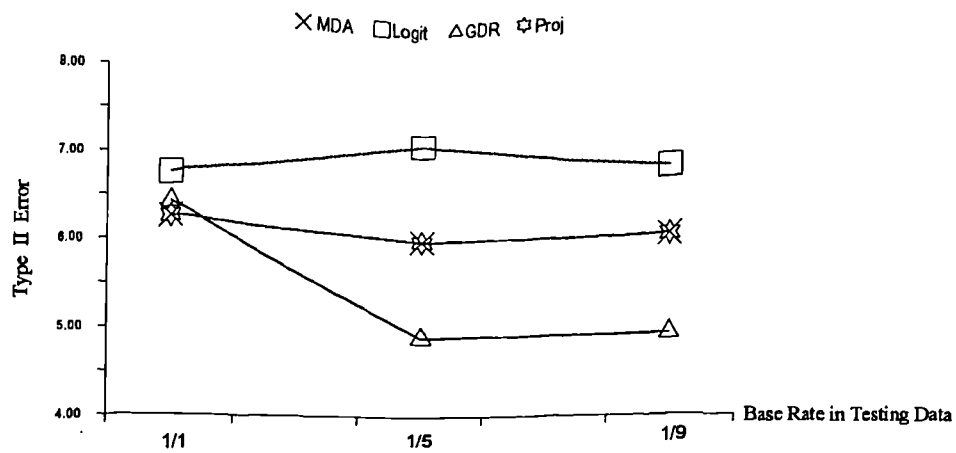


Figure 10.6.17 Type II Error vs. Variuos Base Rates of Testing Data when Base Rate of Training Data is Fixed to 1/1 Using Testing Data Results

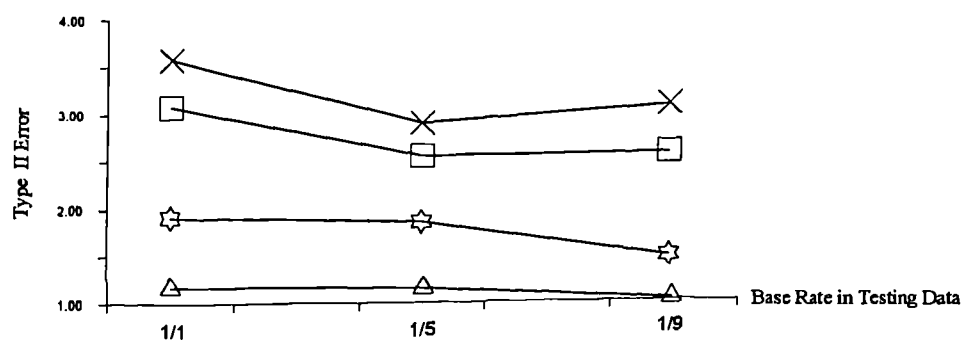


Figure 10.6.18 Type II Error vs. Various Base Rates of Testing Data when Base Rate of Training Data is Fixed to 1/5 Using Testing Data Results

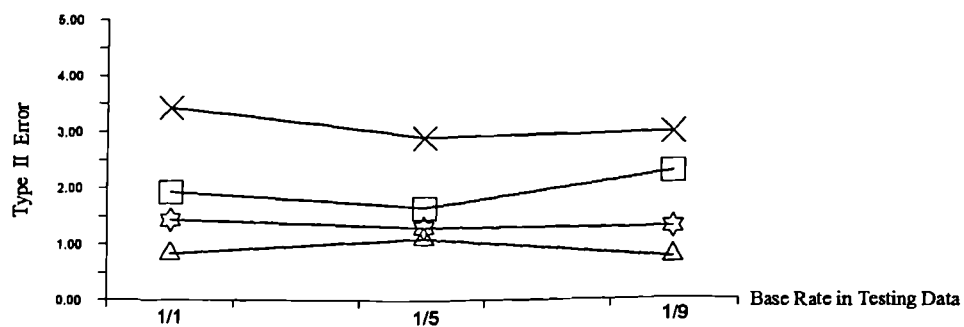


Figure 10.6.19 Type II Error vs. Various Base Rates of Testing Data when Base Rate of Training Data is Fixed to 1/9 Using Testing Data Results

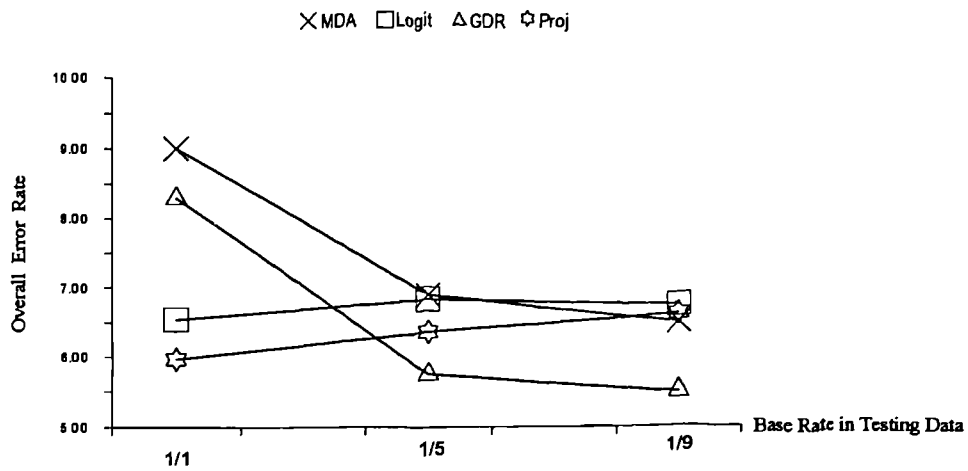


Figure 10.6.20 Overall Error vs. Various Base Rates of Testing Data when Base Rate of Training Data is Fixed to 1/1 Using Testing Data Results

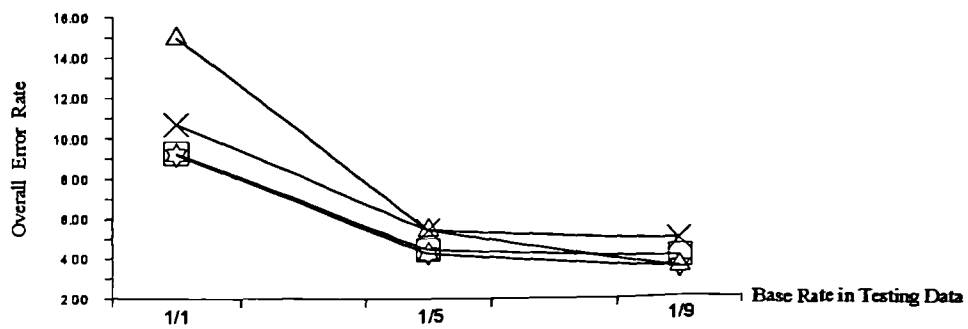


Figure 10.6.21 Overall Error vs. Various Base Rates of Testing Data when Base Rate of Training Data is Fixed to 1/5 Using Testing Data Results

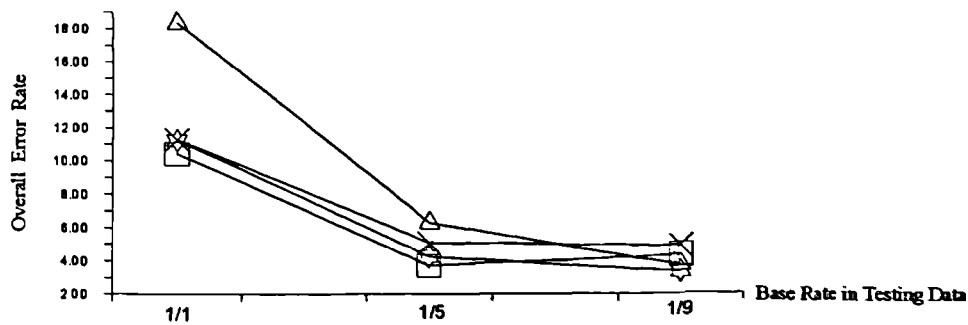


Figure 10.6.22 Overall Error vs. Various Base Rates of Testing Data when Base Rate of Training Data is Fixed to 1/9 Using Testing Data Results

that the Type I and Type II error rates are judged most balanced in the estimation process in spite of the imbalanced proportion of the two groups in the sample. This outcome is probably due to the fact that its classification rule is based on conditional probability in place of composite score of the MDA or composite predicted value of the ANNs.

## **10.7 Summary and Conclusions**

This chapter has provided several insights concerning financial distress prediction models for a real financial data. First of all, an empirical comparison between two conventional statistical methods, MDA & Logit, and two neural network approaches, GDR & Proj was conducted. From the results of this experiment, it is apparent that for the bankruptcy prediction problem, neural networks offer a viable alternative approach. Nonetheless, one caution in this approach is that ANNs have a potential to deteriorate the generalisation ability because they tend to overemphasise classification accuracy instead of recognising the main pattern in the learning phase.

Secondly, with respect to the impact of training data size on predictive ability, neural networks achieve a better classification performance than do statistical methods, yet this advantage decreases as the sample size increases. Further, neural networks provide more stable predictive power since the standard of deviation of misclassification rates on 20 replications is smaller than that in statistical methods. But this stability is less in a small sample size than in a large sample size. Basically, ANNs are more robust than neural networks to the sample size.

Thirdly, the study assessed the influences of a choice-based sample bias. This bias leads to asymptotically biased probability estimates. In the experiment, six choice-based samples designed to induce increasing amounts of bias were examined for their error rates by comparing unadjusted and adjusted procedure (WCOP) over four discriminating techniques. The results clearly demonstrated the existence of a bias for choice-based samples when the unadjusted procedure is used, decreases in the bias as the sample composition approaches the population composition, and the elimination of the bias using the adjustment method. However, the bias does not, on average, affect the overall misclassification rates for all four methods.



Fourthly, the investigation of the sensitivity of optimal cutoff points to the misclassification costs of Type I and Type II errors in the business failure prediction was performed. From the results we can suggest that bankruptcy prediction models are generally applicable over a wide range of possible misclassification costs since the optimal cutoff points are broadly robust to different relative misclassification costs regardless of which discriminating technique is used.

Finally, we test whether the predictive accuracy is significantly affected by the different proportion of nonbankrupt firms to bankrupt firms (base rate) between training and testing sets for four methods, and if any particular method is less affected when this difference occurs. Results have indicated that for both statistical and ANN approaches, they all provide the best performance when the base rates of training and testing are equal. Further, the closer the base rate between training and testing, the greater the accuracy for both statistical and ANN approaches. Generally speaking, the four methods were shown perform well in predicting nonbankrupt firms. However, the statistical methods have better performance in predicting bankruptcy as the proportion of nonbankrupt to bankrupt firms diverges. In terms of generalisation ability, since the statistical methods, particularly in Logit, are less affected by the different composition of the two groups between training and testing, they are judged to be more robust than neural networks.

In the final chapter the results of both the simulation study and the empirical study will be summarised. The theoretical and practical contributions and implications will be discussed. The strengths and limitations of this research will also be presented. Finally, the directions for future research are considered.

## Chapter Eleven

### OVERALL COMMENTS AND DIRECTIONS FOR FUTURE RESEARCH

#### 11.1 Introduction

In the concluding chapter, it is appropriate to summarise the main conclusions and the results of the previous ten chapters, and suggest directions for future research.

This chapter is set out in five further sections. The next section presents the new elements in this thesis. The second section discusses the new findings or the results which conform or contradict existing literature; as well as likely reasons of the contradictions and the implications of the new findings obtained. The recommended discriminator identified for use in suitable data conditions is described in the third section. The inherent limitations of this research are reviewed in the fourth section. The last section considers some possible extensions which may provide a basis for future research in this area.

#### 11.2 The New Elements in this Thesis

The new elements in this thesis include that

1. The development of the models has made use of simulation and of real data. Particular attention has been paid to the type of error (type I or type II) as well as the overall accuracy. An exploration of the situations in which one technique dominates the others has discovered that the Projection network performs best when the data most departs from normality (high skewness and in the presence of outliers).

Based on the results of this study, Projection net undoubtedly opens a new window on classification problems and offers good prospects for future elaboration.

2. The bias in factor analysis which has been overlooked for a long time was clearly pointed out by our examples. We thus suggest the discard of this procedure in the selection of predictors if the classification accuracy is the primary concern. (section 5.4 and 5.5). It

is hoped that subsequent studies can avoid the problem of losing valuable information for discrimination, when coping with the need to reduce the set of independent variables.

3. An approach to eliminate the choice-based sampling bias has been proposed. The approach required neither the assumptions of independent variables (predictor variables) nor the construction of maximum likelihood functions. It can thus be applied to any discriminating technique. This solution, which uses the weighted cutoff point (WCOP), on one hand, allows the matched sample design controlling over industrial or size effect; on the other hand, it removes the choice-based sampling bias. (section 8.3 and 10.4)
4. This thesis has investigated the impact of some of the factors on models of failure prediction: the role of the sample size, the effect on optimal cutoff to various misclassification costs, and the intertemporal and out of sample ability of the models (section 10.6 and 10.7). These factors affect both statistical models and artificial neural network models.

### **11.3 The Findings, Explanations and Implications**

The findings which are new and the results which conform or contradict other studies are summarised in Table 11.3.1. Our understanding about the issues, the explanations why these contradictions may occur and the implications of new findings are given below.

First of all, the arguments and explanations for the contradictions are developed. For item 4 in Table 11.3.1, there is a question whether it truly indicates the inferiority of Logit. While the overall performance of logistic regression was shown the worst in terms of the total number of higher error rates, none of the differences between Logit and MDA were found significant at the 95% level. This is displayed in Figure 11.3.1 for training samples, and in Figure 11.3.2 for testing samples, respectively. Moreover, Logit is barely inferior to the GDR in most of the cases as presented in the figures. In particular, 26 of 36 testing results in Logit have shown lower error rates than those in MDA. In this regard, we could say that Logit produced better generalisation compared to MDA.

Table 11.3.1 Summary of the Findings in this Thesis

Results	Confirm	Contradict	New Findings	Relevant Sections or Pages of the Results in this Thesis	Where the Results Confirm or Contradict	Comments
1. The pattern of errors between ANNs and STMs is quite different			X	section 7.3.1 and 7.4.1		According to our results, MDA and Logit produce the same pattern of Type I and Type II error rates. By contrast, GDR and Projection neural networks always provide identical direction of the two type errors.
2. ANNs perform at least as well as STMs	X			section 7.3.1 and 7.4.1	[Gallinari et al., 1988]; [Dutta and Shekhar, 1988]; [Salchenberger, Cinar and Lash, 1992]; [Kim, Weistroffer and Redmond, 1993]; [Wilson and Sharda, 1994]; [Altaman, Marco and Varetto, 1993]; [Wilson and Sharda, 1994] etc.	Except for case Aa1, Aa2, Ab1 and Ab2, which have shown no significant differences between ANNs and STMs, ANNs always performs better than STMs across all other data conditions.
3. Projection overall offers the best performance among four methods			X	section 7.3.1 and 7.4.1		The introduction of Projection is novel in this area.
4. Logit generally produces the worst classification among four methods in terms of total number of higher mean error rates.		X		section 7.3.1 and 7.4.1	[Bell, Ribar and Verchio, 1990]; [Halperin, Blackwelder and Verter, 1971], [Press and Wilson, 1978], [O'Hara et al., 1982], [Hosmer et al., 1983a, 1983b]	In fact, there are no significant differences between Logit and MDA for all cases; and Logit is barely inferior to GDR in most situations at the 95% confidence interval level. In addition, Logit offers better generalisation ability than MDA does.
5. The classification ability in Projection is worse for equal group dispersion than for unequal one			X	Page 163, 164, 184 and 185		We try to explain this curious outcome by plotting the discriminant functions of four methods on relevant cases. We believe that it is connected with the way Projection algorithm forms its decision boundary.
6. ANNs are more robust to outliers than STMs	X			section 7.3.5 and 7.4.5	[Denton et al., 1990]; [Lippmann, 1987]	This phenomenon is obvious in all cases with C.

**Table 11.3.1 Summary of the Findings in this Thesis**  
(continue)

Results	Confirm	Contradict	New Findings	Relevant Sections or Pages of the Results in this Thesis	Where the Results Confirm or Contradict	Comments
7. ANNs, especially in GDR, have the tendency to produce higher Type I error than STMs			X	section 7.3.2 and 7.4.2		From our analyses, worse Type I performance frequently occurs in normal distribution or the data with outliers.
8. ANNs are more capable of capturing complex decision frontiers than STMs	X			section 7.3.5 and 7.4.5	[Gallinari et al., 1988]; [Lippmann and Beckman, 1989]; [Denton et al., 1990]; [Lippmann, 1987]	It is evident when the observations of two groups strongly overlap. This is related to ANNs' capacity to cope with nonlinearity.
9. ANNs show better learning but worse generalisation	X			section 7.3.5 and 7.4.5	[Tam and Kiang, 1992]; [Berry and Trigueiros, 1993]	This is known as overfitting problem.
10. There are good learning but bad generalisation in small size no matter what technique is used	X			section 10.3.1	[Sietsma and Dow, 1991]; [White, 1993]	Small sample size is not recommended for the purpose of prediction.
11. For large sample size, ANNs clearly shows its superiority. However, this advantage decreases as the sample size increases.			X	section 10.3.1		It is no doubt that ANNs require a lot of data because they are empirical models. Nevertheless, good data is more important than large data.
12. The performance of ANNs in small sample size is less stable than that in large sample size.			X	section 10.3.1		Small sample size may not suffice for detecting the main pattern. Overfitting problem also confuses the performances.
13. There exists a choice-based sampling bias in STMs	X			section 10.4.1	[Manski and Lerman, 1977]; [Zmijewski, 1984]; [Dopouch, Holthausen and Leftwich, 1987]	There is overclassification of bankrupt firms, and a underclassification for nonbankrupt firms. Meanwhile the error rates have functional relationship with the decreasing sample.

**Table 11.3.1 Summary of the Findings in this Thesis**

(continue)

<b>Results</b>	<b>Confirm</b>	<b>Contradict</b>	<b>New Findings</b>	<b>Relevant Sections or Pages of the Results in this Thesis</b>	<b>Where the Results Confirm or Contradict</b>	<b>Comments</b>
14. There exists a choice-based sampling bias in ANNs			X	section 10.4.1		The results are similar to those in STMs. But no other study on ANNs has been conducted yet.
15. The WCOP approach can eliminate most, if not all, the choice-based bias present in four methods.			X	section 10.4.1		This procedure can be applied by different discriminating methods and avoids the complicated calculation produced by WESML.
16. Bankruptcy prediction models are generally applicable in STMs as well as in ANNs when the cost of Type I error is much higher than that of Type II error.	X			section 10.5.1, 10.5.2, 10.5.3 and 10.5.4	[Koh, 1992]	This implies that an inaccurate estimation of Type I and Type II error costs is not a very serious limitation in bankruptcy prediction within such range of error cost ratios.
17. The optimal cutoff points are relative sensitive within the low cost ratio area.		X		section 10.5.1, 10.5.2, 10.5.3 and 10.5.4	[Koh, 1992]	This phenomenon is especially obvious in MDA.
18. The best performance usually occurs when the base rate between learning and testing sample is equal (i.e. 1/1 vs. 1/1, 1/5 vs. 1/5, and 1/9 vs. 1/9) for all four methods.	X			page 294	[Wilson and Sharda, 1994]	This means, equal proportion of bankrupt and solvent firms between training and testing samples is needed if high predictive ability is required.
19. The 1/1 case does not necessarily provide better prediction power than 1/5 or 1/9 cases on any of these four approaches.			X	page 294, 295	[Wilson and Sharda, 1994]	It implies that a one-to-one matching design is not the best choice if minimising the overall error rates is the major concern.
20. STMs have better performances in predicting bankruptcy as the proportion of bankrupt to nonbankrupt firms diverges.	X			page 296	[Wilson and Sharda, 1994]	When the data of bankrupt companies is not enough, STMs is more suitable for prediction.

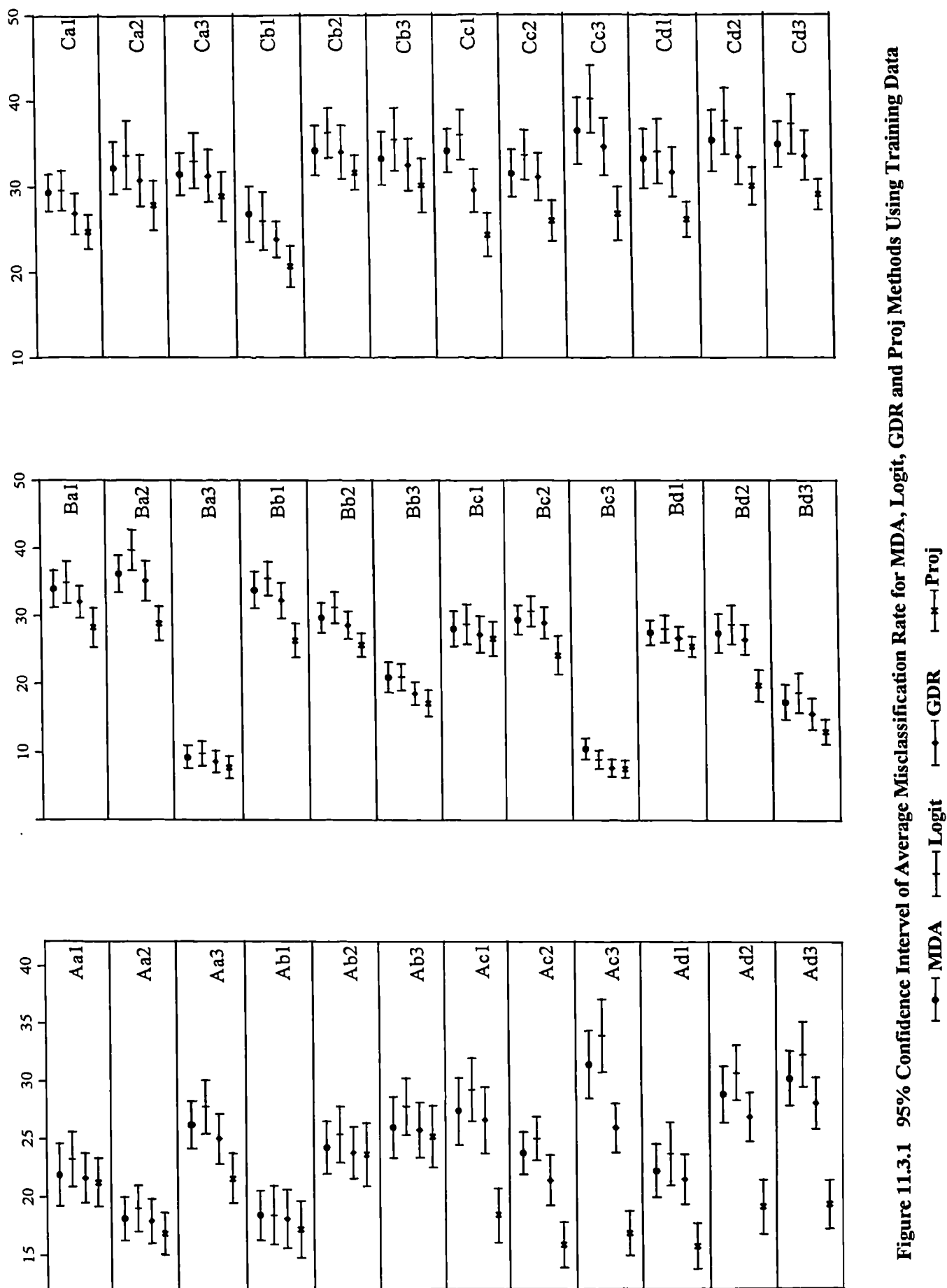


Figure 11.3.1 95% Confidence Interval of Average Misclassification Rate for MDA, Logit, GDR and Proj Methods Using Training Data

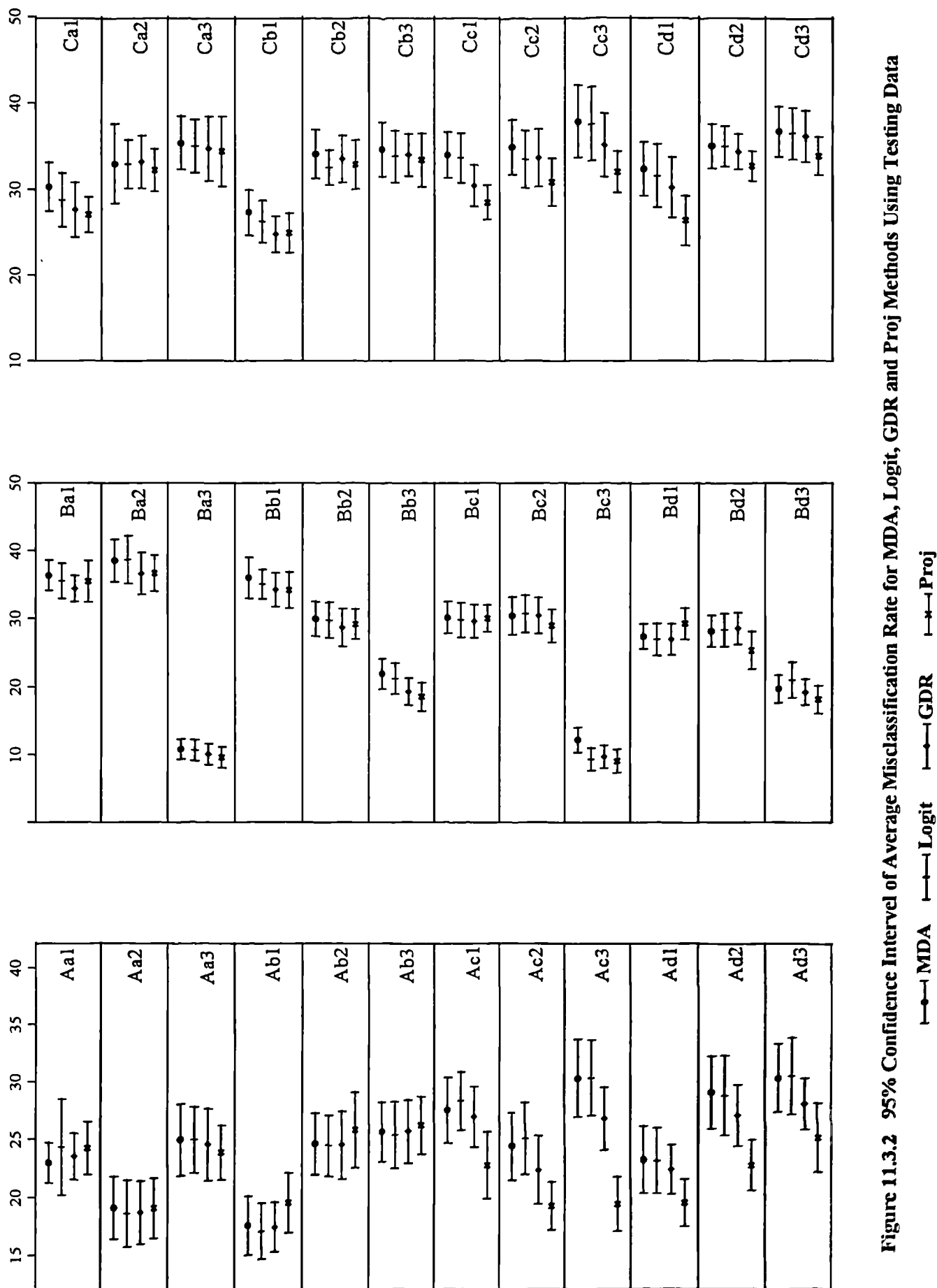


Figure 11.3.2 95% Confidence Interval of Average Misclassification Rate for MDA, Logit, GDR and Proj Methods Using Testing Data



One interpretation of the unfavorable learning results of Logit is that the linear combination of variables in multivariate logistic formulation is not always an appropriate model [Gordon, 1974], since some of interaction effects, which were proved significant in our data, may not be expressible.

With regard to the finding of item 17, the optimal cutoff points are relatively sensitive within the low cost ratio range. It is somewhat inconsistent with Koh's study [1992], which concluded that optimal cutoff points are robust to a broad range of misclassification cost ratios in the going-concern prediction context. The main cause for this inconsistency is in the different estimation procedures used. Koh's model incorporates the prior probabilities into his Logit model by adjusting the constant term as suggested by Maddala [1988]. This changes the shape and the degree of separation of posterior distributions for two groups. However, our WCOP procedure instead changes the location of the optimal cutoff points instead in response to the change of the ratio of costs. The idea is presented in Figure 11.3.7.

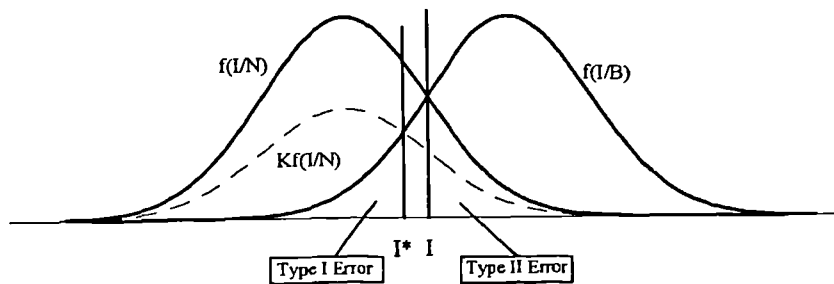


Figure 11.3.7 Change of Optimal Cutoff Point to the Ratio of  $C_{II}$  to  $C_I$

As can be seen, when the shape of  $f(I|B)$  or/and  $f(I|N)$  have a steep side (vertical side), there would be no change in cutoff  $I^*$  as  $K$  (the ratios of costs) increased. A very sensitive  $f(I|B)$  or/and  $f(I|N)$  would be one in which the side was almost horizontal. The sensitivity of the cutoff depends on the gradient at the intersection. Usually steep gradients occurs at the middle of the distribution while flat gradients occurs at the tail of the distribution. Thus when the cost ratio is high, and the intersection is at the tails, the cutoff would be more sensitive for changing  $K$  than for low cost ratios.

This explanation can be expressed mathematically. The cutoff point is the root of equation (11.3.1)

$$K f(I|N) - f(I|B) = 0 \quad (11.3.1)$$

To see the sensitivity of the cutoff point to changing cost ratios, take total differentiation

$$[dK f(I|N) + K f'(I|N) dI] - f'(I|B) dI = 0 \quad (11.3.2)$$

then solve for  $dI/dK$ , one has

$$\frac{dI}{dK} = -\frac{f(I|N)}{K f'(I|N) - f'(I|B)} \quad (11.3.3)$$

One can see from this result that the cutoff depends on  $f'(I|N)$  and  $f'(I|B)$ . To demonstrate further, if both  $f(I|B)$  and  $f(I|N)$  are assumed normal distributions with means  $\mu_B$  and  $\mu_N$ , and standard deviations  $\sigma_B$  and  $\sigma_N$ , respectively. The equation (11.3.3) can then be written as

$$\frac{dI}{dK} = \frac{\sigma_N \sigma_B^2 \text{Exp}\left[-\frac{(I-\mu_N)^2}{2\sigma_N^2}\right]}{K (I-\mu_N) \sigma_B^2 \text{Exp}\left[-\frac{(I-\mu_N)^2}{2\sigma_N^2}\right] - (I-\mu_B) \sigma_N^2 \text{Exp}\left[-\frac{(I-\mu_B)^2}{2\sigma_B^2}\right]} \quad (11.3.4)$$

It can be seen easily from (11.3.4) that as  $\sigma_B \rightarrow 0$  or  $\sigma_N \rightarrow 0$ , then  $dI/dK = 0$ . It implies that  $I^*$  will not change with  $K$  in this case. On the other hand, if  $\sigma_B \rightarrow \infty$ ,  $dI/dK = \sigma_N/K(I-\mu_N)$ , when  $\sigma_N = 0$ ,  $dI/dK = 0$ ; when  $\sigma_N \rightarrow \infty$ ,  $dI/dK \rightarrow \infty$ , which also proves the above comment.

The distributions of  $f(I|B)$  and  $f(I|N)$  for four methods in our real data was shown in Figure 10.2.4 to 10.2.11 (page 258, 259), it indicates that most of frequency distributions are either in flat (horizontal) shapes at the center or in the steep (vertical) shapes at the tail. Hence it confirms the explanation that the optimal cutoff points are relatively sensitive within the low cost ratio than high cost ratio.

Except for the comments in Table 11.3.1, which has already discussed the views of new findings, some important implications are stressed here.

The pattern of errors between ANNs and STMs has been shown different. On the other hand, one surprising outcome in this study is that equal variance-covariance structure had an adverse effect on Projection algorithm. We suspect that these results were connected with the different ways the four methods formed their decision boundaries. Plots of data for case Aa1 Ab1, Ac1 and Ad1 (relevant and representative cases) showing the discriminant

function generated by each method try to explain the possible causes (Figure 11.3.3 to 11.3.6).

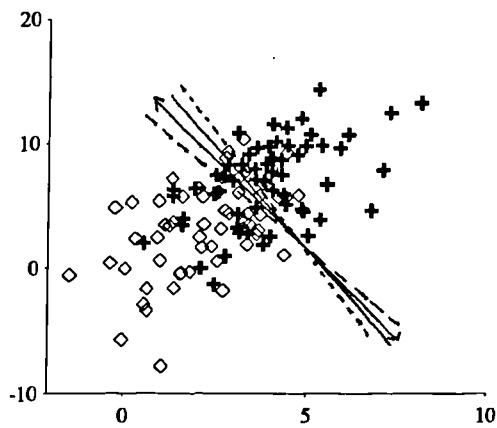


Figure 11.3.3 Discriminant Functions on Aa1 Data

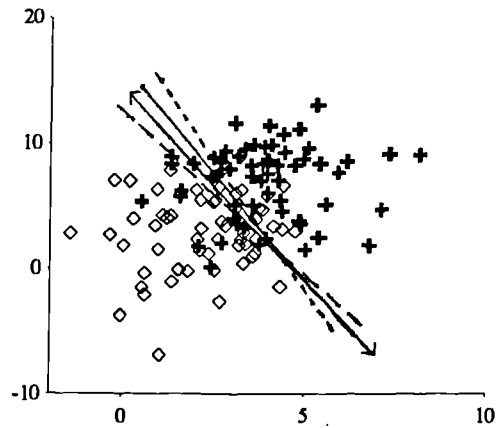


Figure 11.3.4 Discriminant Functions on Ab1 Data

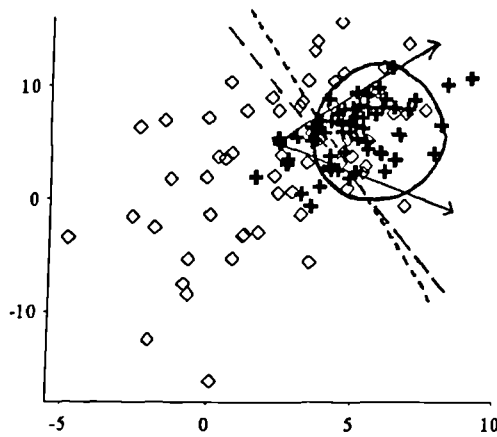


Figure 11.3.5 Discriminant Functions on Ac1 Data

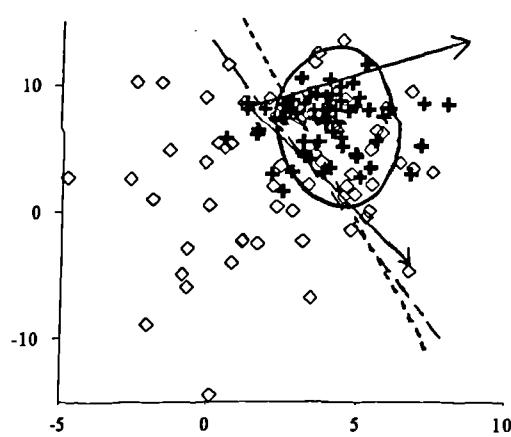


Figure 11.3.6 Discriminant Functions on Ad1 Data

-----MDA      ←→GDR  
 ---Logit      —Proj

It is observed that the position of discriminating functions in STMs are rather sensitive to group dispersion structure, while GDR and Projection that form their boundaries simply target the minimisation of error counts. In other words, the classification rule of ANNs does not seem to take covariance matrices effects into consideration. In particular, the philosophy of Projection immediately places its closed boundary around input points, then it minimise the output error through gradient descent (the optimal solution could be closed or/and open boundaries). This computation process had nothing to do with any data assumption. In our study it is difficult to assure what really causes this curious phenomenon as it emerged from analysis simultaneously conducted on three factors. This should be

investigated further to determine whether the above surprising outcome is a coincidence or can be generalised.

Regarding the unstable performance of ANNs in small sample size, our observation is that random subset of the whole data may generate insufficient information representing the main pattern, thus it strongly affected the performance. This should serve as a warning: if data is not enough, reconsider using a neural network. On the other hand, large sample size has the chance to complicate the tasks of dissection and evaluation so as to deteriorate the performance. This is also a cause for an alarm: the key issue may not be the technique itself, but the underlying structure in the available data. That is, if data is not good enough, reconsider using any of the methods.

The composition of Type I and Type II errors is a significant topic in this thesis. The costs of errors and the decision context are important in developing the models. While previous studies have emphasised overall accuracy, we approached the comparisons that were made from a new angle. In this research, evidence indicated an improvement in overall performance in ANNs where Type II errors were involved. In many cases Type I errors conversely deteriorated, especially in GDR. Accordingly ANNs' outperformance may be conditional.

However the Projection network generally exhibited its superiority not only quantitatively (i.e., overall accuracy), but also qualitatively (lower Type I and Type II error). Hence, if ANN is used as a discriminating tool in predicting bankruptcy, a GDR is not to be recommended, but the Projection is a preferable method whatever the Type I or Type II error is more costly.

In view of our and Koh's results [1992] concerning the general insensitivity to variation in relative costs, the question now lies in the matter of the desirability of such insensitivity to a model. Superficially speaking, this means that bankruptcy prediction are generally applicable over a wide range of possible misclassification costs. However, this could also mean that the model fails to reflect the significance of this information. Thus, banks and auditors, or whoever is concerned about this issue, are unable to make a correct judgment based on their own interests.

On the other hand, our findings, which revealed the relative sensitivity of optimal cutoff points to closer ratios of two type error costs, told us that nonconsideration of asymmetric loss function produces some degree of risk of nonoptimal decision in such cost ratio area. Further, a deeper significance of this phenomenon in MDA is believed to be linked to the

nature of composite score of LDF rather than the conditional probability of Logit or predicted value of ANNs, which were obtained through the mapping and transformation of sigmoid function.

The tendency of obtaining higher error rates on smaller sample proportion group in ANNs compared to STMs is an interesting finding. It resulted from their underlying logic of caring too much about overall accuracy, too little concern was given to inherent implication of the data, thus the misclassification on the group of rare occurrence is amplified.

One other interesting result is that in a situation where the proportion of bankruptcy to nonbankruptcy between training and testing samples is the same (such as 1/1 vs. 1/1, 1/5 vs. 1/5 and 1/9 vs. 1/9), the predictive ability of 1/1 composition case is not necessarily better than those of 1/5 or 1/9 composition ones. The implication may be that the one-to-one matching is not preferable to the one-to-many if match criterion is chosen.

Furthermore, one very important difference between ANNs and STMs that needs to be addressed is that ANNs are basically a dynamic process. ANNs can continue to evolve as new cases are added to the network, this is vastly different from the static batch updating in statistical methods. This nature enables ANNs to cope with a changing environment over time by adapting gradually to new cases representing changes in the model. Although our research has not covered this issue, we remain aware of it. We believe that this fundamental difference has a great impact on ANNs' future development and their applications.

Finally, some comments about STMs and ANNs are made. In one sense neural networks are little more than nonlinear regression and allied optimisation methods. However, they do have a methodology of their own. We must admit that, on one hand, their success is a warning to statisticians who have worked in a simply structured linear world for too long. However, on the other hand, the experience of statistician in modelling, in assessing competing models can and should be brought to bear in what is now the discipline of ANNs. The value of model generalisation is one of such ideas. Published research has talked about this issue a lot. But we still feel that it is very important to remind the user that the better performance of ANNs is likely to be a veiled case of overfitting. We remain convinced that in most real-world problems, there is other knowledge which can be used to guide modelling and so enables much more generalisation.

Statistics, is more difficult sometimes, but that is also interesting and challenging. Like Aharonian [1992] commented ANNs on financial applications that neural networks may be cold-fusion of statistics world, and so people are going back to learning and finding statistics.

#### **11.4 Data Conditions and Recommended Discriminating Techniques**

One of the important findings of this research is the ability to identify problem and input characteristics when one discriminating approach provides better performance than others. Taking all aspects into account, the suitable techniques for which problem characteristics to be used, and under what circumstances are summarised in Table 11.4.1.

#### **11.5 Limitations of the Study**

There are some limitations to the study. Firstly, due to an intent to facilitate the interpretation of the results, the cases in the simulation study were limited to bivariate populations when reasonable comprehensive testing conditions were being promoted. Although it provided valuable insights into the strengths and weaknesses of each of the four techniques, however, our ability to generalise results to the  $n$ -dimensional case needs to be carefully explored.

Secondly, the predictor variables in our empirical study were selected on the basis of the two theories of bankruptcy developed by Hudson [1986] and Laitinen [1991], which incorporated only 12 financial ratios. No matter which technique is used, the selection of financial ratios as independent variables in predicting business failure is always a major problem. Although these twelve financial ratios used in distinguishing failure from nonfailure tended to be key factors in the present study, seeking the optimal set of financial indicators for bankruptcy prediction should always be the researcher's chief goal.

Thirdly, our empirical study was constructed with and validated on Datastream companies only. Although the underlying characteristics of bankruptcy and nonbankruptcy may be the same whether they are in Datastream companies or not, the effectiveness of the results should be tested on other real financial data or other countries' situations in order to obtain more reliable conclusions.

Another limitation is that the empirical experiment has concentrated only on one-year ahead prediction. The effectiveness of ANNs and their comparison with STMs more than one year prior to the date of bankruptcy have not been evaluated. It would be necessary to investigate the performances and make a comparison among these four methods on a long term as well as on a short term basis.

**Table 11.4.1 Data Conditions and Recommended Discriminating Methods**

<b>Data Conditions Favourable to Artificial Neural Networks</b>	<b>Data Conditions Favourable to Statistical Methods</b>
ANNs can be used in all data situations for bankruptcy prediction if the overall classification accuracy is the sole objective. The Projection method is particularly preferable when lower Type I error is required.	STMs can be effectively used to the data if departure from normality is not severe or when group dispersions are not strongly heterogeneous. MDA is especially recommended for this situation.
ANNs are strongly recommended in the presence of extreme values and noisy data.	STMs are better methods for normal model when misclassification cost of Type I errors is much higher.
ANNs are able to find the pattern when the underlying characteristics of data or function is totally unknown.	STMs are suitable for data with enormous observations and huge independent variables when the pattern is not complicated and running time is also an important factor.
ANNs have the advantage of running adaptively to cope with dramatic change of environment over time.	STMs has less tendency of overfitting and can efficiently extract key features for data with stable underlying relationships over time.
ANNs provide more flexibility in locating the best model, allowing for sophisticated model builders to have a wide of model parameters and properties for fine-tuning models	STMs are simple to develop and have an advantage of transparent interpretability.
ANNs are better developed using a reasonably large sample size.	Logit is a better choice for classifying when the proportion of failing firms is rather small, and the misclassification cost of Type I errors is relatively high.

## 11.6 Directions for Future Research

The limitation identified above suggest possible directions for future research. For example the first limitation highlights the possibility of constructing a multivariate population model. For more complicated data conditions, future work can concentrate on certain specific cases in order to promote deeper comparisons and robustness tests. For instance, in our simulation study, the contradictory result about classification performance on equal and unequal variance-covariance matrices data conditions between STMs and ANNs is believed to be a suitable subject for future study when this factor is solely investigated and the number of variates is allowed to change.

As mentioned before, finding an optimal set of predictor variables is an important goal in bankruptcy prediction. Factor analysis is one of these techniques. It is often used to create a reduced and best predictor set. However, it has been demonstrated to have the risk of losing valuable information needed to distinguish one group from another. There is the need for research to explore other directions and possibilities. One possible solution relies on the learning capacity of the neural network. Since searching for clusters is one of the key pattern recognition techniques which is of central importance in neural networks, the ability of neural networks to extract important factors is one of their major functions [Hertz, Krogh and Palmer, 1991]. Anderson, one of the leading researchers in the field of neural networks writes: "It has become clear in the past couple of years that some older techniques from pattern recognition and statistics are similar to some of the ideas developed for neural networks." [Anderson et al., 1990]. He mentioned in particular that an idea from statistics that is coming to greater and greater prominence in neural networks is the multivariate statistics theory of principal components [Bharath and Drosen, 1994]. The ability of ANNs to elicit key factors leads us to speculate that, in contrast to the conventional habit of using factor analysis to determine the appropriate independent variables for a neural network, the reverse may be worth trying especially when no prior theory guides the model formulation [Trigueiros and Taffler, 1995].

One way to achieve this goal is by applying machine learning techniques such as the genetic algorithm. Back et al. [1995a] have used this algorithm to successfully find which financial indicators are the best bankruptcy predictors. The genetic algorithm picked out 20 financial indicators from the total data set of 54 indicators. This reduced variable set provided



"extremely good" classification accuracy results. Other research undertaken by Levitt [1995] applied the modified genetic algorithm to foreign exchange trading in order to predict the sign of the returns of the USD/DEM exchange rate. The subset of indicators obtained from the modified genetic algorithm resulted in more accurate prediction performance than an unprocessed universe of indicators.

Very interesting research conducted by Berry and Trigueiros [1993] has aimed to enhance the interpretability of neural network parameters in the account context so that they help an analyst to search for appropriate ratios before model building. Instead of financial ratios themselves, eight accounting variables extracted from eighteen financial ratios were used as input variables. Based on the assumptions of many accounting variables with cross-section distributions that are approximately log normal, the logarithmic transformation is thus applied. If the linear combination of this transformed form is reversed, it will appear to be equivalent to a complex ratio form. That is, the linear combination of an observation in log space is equivalent to ratio form in ordinary space. They explained that "if the values input to an MLP are the logs of variables, then the neurones in the first hidden produce NETs that represent complex ratios." (p.110). The underlying idea here is that the weights connecting input variables with the hidden layer's nodes seem to be the exponents of the extended ratios involved in the optimal solution. In this application, forming ratios in hidden nodes is very different from other studies and provides a new view on this issue.

Other useful direction is to investigate how far in advance and how well ANNs can effectively predict the financial health status of companies. Such information is helpful to management, investors, creditors, regulatory agencies, and others who are interested in the long-term financial condition of companies. Consequently, an immediate extension to more than one year prior to failure should be considered.

The following several points can also be considered as possible extensions

1. The determination of optimal topology in ANNs is often the model builder's main concern. This concern is especially for the structure of the hidden layer. It is hidden layers and nodes that make MLP a powerful tool. The outputs of hidden layer nodes can be considered as new variables, which can themselves contain interesting information about the relationship being modelled. Therefore, the development of a formal procedure or guideline for determining the number of hidden layers and number of hidden nodes is

essential. For some complex problems, Surkan and Singleton [1992] have demonstrated that a redistribution of the nodes from one to a pair of hidden layers can improve classification accuracy, and a significant advantage arose even when the layers were ordered so that a smaller number of hidden nodes received their inputs directly from the input variables. The impact of structural variations in the hidden layer on predictive ability is believed to be an important subject in ANNs for future investigation.

2. Some nonparametric discriminant procedures based on linear programming (LP) such as MSD (minimise sum of deviation) and LAD (least absolute deviation), have been shown to be competitive discriminating tools comparing favourably with Fisher's LDF under certain circumstances [Freed and Glover, 1986]; [Bajgier and Hill, 1982]; [Joachimsthaler and Stam, 1988]; [Stam and Jones, 1990]; and [Lee and Ord, 1990]. Proponents of the LP methods have cited several advantages in these formulations: (1) The methods are independent of classical statistical theory; there need be no assumptions regarding multivariate normality or equality of variance-covariance matrices and little formal inferential theory. (2) The ability to perform LP sensitivity analysis and effective classification is an advantage that outweighs the loss of any statistical information. Since ANNs are also nonparametric procedures, the comparison between LP models and ANNs is recommended for investigation in future work.
  
3. Neither this research nor previous studies have included qualitative variables in predictors when comparing STMs to ANNs. The effect of incorporating qualitative variables such as the type of industry, firm size and macro-economic considerations in ANNs for predicting failure has not been fully explored. Many authors using statistical methods have endeavoured to employ these factors in order to create more stable models. Altman [1973] and Mason and Harris [1979] attempted to establish industry specific models. Izzan [1984] and Platt and Platt [1990] employed industry relative ratios to cope with this instability problem. Dambolena and Khoury [1980]; Rose, Andrew and Giroux [1982]; and Mensah [1984] included an estimation of business cycles in order to unveil macro-economic influences. The evidence revealed that they reduced the heterogeneity of firm data, provided more stability and offered better predictive ability, although these qualitative variables apparently violate the multivariate normality assumption of the MDA

method. ANNs, free of underlying parametric assumptions that statistical approaches require, are more suitable for utilising these qualitative indicators as predictors.

4. Previous corporate failure prediction models have focused mainly on the conventional failing/nonfailing dichotomy. A more complex model would be much more useful for decision makers. For example, a company's assessment which falls into the middle area can be seen as a gray zone and needs to be subjected to closer evaluation. This would be more consistent with the practical operations of financial institutions when evaluating the granting of a bank loan. In effect, multiple classification problems also involve bond rating as part of this subject. Corporations often raise additional cash by issuing bonds. These corporate bonds are essentially loans at a given rate of interest and must be paid off within a specified time period. Potential investors decide whether or not to invest in the bond based on the rating by the rating agencies—Moody's or Standard & Poor's. The quality ratings (Aaa, Aa1, Aa2, Aa3, and so on) strongly affect the interest rate the corporation will pay. The lower rated bond will lead to the offer of a higher rate than a higher rated bond which attracts investors because of the implied higher risk. Thus, bond rating is crucial both for corporations and investors. Previous bond rating research largely used statistical tools, primarily regression analysis [Horrigan, 1966]; [West, 1970], logistic analysis [Ederington, 1985]; [Kaplan and Urwitz, 1979], and discriminant analysis [Belkaoui, 1980]; [Pinches and Mingo, 1973] which are very similar to bankruptcy prediction techniques. Since multi-state prediction is more frequently used and more meaningful in practice, future work in ANNs needs to focus on this subject.
5. Most research dealing with the application of ANNs to financial forecast has made use of a feedforward layered neural network together with the backpropagation training algorithm. However, there has been some, though little, work devoted to other ANN models in accounting, especially for bankruptcy predictions. Serrano Cinca et al. [1993] carried out a study to predict Spanish bank failure using a self-organising feature map (SOFM). SOFM [Kohonen, 1989, 1990] is an unsupervised neural network and is also known as the Kohonen map. Researchers claim that they have achieved better results with the self-organising maps than with the backpropagation nets. Back et al. [1995b] also conducted a comparative analysis on bankruptcy prediction performance using the

Kohonen net, the backpropagation net, and the mean field annealing algorithm (Boltzmann Machine). Their results indicate that the backpropagation net performed better than the other two methods. These results contradict those in the study by Serrano Cinca et al. [1993]. However, no matter which method is ultimately the best, we have found that apart from backpropagation algorithm, there exist other ANN models such as SOFM, Boltzmann Machine, or the LVQ (Learning Vector Quantization Networks) [Kohonen, 1989, 1990] and the fuzzy neural network [Wong et al., 1992], etc. which could be very interesting tools for the analysis and forecasting of financial data, and which are thus worthy of future elaboration.

## BIBLIOGRAPHY

- Ahalt, S. C., Garber, F. D., Jouny, I. and Krishnamurthy, A. K. [1989] "Performance of Synthetic Neural Network Classification of Noisy Radar Signal", *Advances in Neural Information Processing Systems*, edited by Touretzky, David. S., Vol. 1, San Mateo, CA: Morgan Kaufmann, 1989, pp.281-288.
- Akaho, S. and Amari, S. [1990] "On the Capacity of Three-Layer Networks", *Proceedings of the International Joint Conference on Neural Networks*, San Diego, Vol. III, 1-6.
- Akaike, H. [1974] "A New Look at the Statistical Model Identification", *IEEE Transactions on Automatic Control*, Vol. AC-19, No. 6, December 1974, pp.716-723.
- Altman, E. I. [1968] "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy", *The Journal of Finance*, September 1968, pp.589-609.
- Altman, E. I., Marco, G. and Varetto, F. [1994] "Corporate Distress Diagnosis: Comparisons Using Linear Discriminant Analysis and Neural Networks (The Italian Experience)", *Journal of Banking and Finance*, 1994, Vol.18, No.3, pp.505-529.
- Altman, E. I. [1973] "Predicting Railroad Bankruptcies in America", *Bell Journal of Economics and Management Science*, Spring, pp.184-211.
- Altman, E. I. [1980] "Commercial Bank Lending: Process, Credit Scoring, and Costs of Errors in Lending", *Journal of Financial and Quantitative Analysis*, November, 1980, pp.813-832.
- Altman, E. I. and McGough, T. P. [1974] "Evaluation of a Company as a Going Concern", *Journal of Accountancy*, December 1974, pp.50-57.
- Altman, E. I. and Spivack, J. [1983] "Predicting Bankruptcy: the Value Line Relative Financial Strength vs. the Zeta Bankruptcy Classification Approach", *Financial Analysts Journal*, November-December, 1983, pp.60-67.
- Altman, E. I. and Eisenbeis, R. A. [1978] "Financial Applications of Discriminant Analysis: A Clarification", *Journal of Financial and Quantitative Analysis* March 1978, pp.185-195.
- Altman, E. I., Haldeman, R. and Narayanan, P. [1977] "ZETA Analysis: A New Model to Identify Bankruptcy Risk of Corporations", *Journal of Banking and Finance*, June 1977, pp.29-54.
- Amemiya, T. and Powell, J. [1983] "A Comparison of the Logit Model and Normal Discriminant Analysis When the Independent Variables are Binary", In: Samuel Karlin, Takeshi Amemiya and Leo A. Goodman. eds., *Studies in Econometrics, Time Series, and Multivariate Statistics* (Academic Press, New York).
- Anderson, J. A. [1972] "Separate Sample Logistic Discrimination", *Biometrika*, 59, pp.19-35.
- Anderson, T. W. [1958] "An Introduction of Multivariate Statistical Analysis", *Technometrics*, May 1964, pp.179-190.
- Anderson, J. A., Pellionsz, A. and Rosenfeld, E. [1990] "Neuralcomputing 2: Directions for Research", Cambridge, MA: MIT Press, 1990.

- Aziz, A. and Lawson, G. H. [1989] "Cash Flow Reporting and Financial Distress Models: Testing of Hypotheses", *Financial Management*, Spring 1989, pp.55-63.
- Baba, K., Enbutu, I. and Yoda, M., [1990] "Explicit Representation of Knowledge Acquired From Plant Historical Data Using Neural Network", *Proceedings of the International Joint Conference on Neural Network*, III155-III160.
- Back, B., Sere, K. and Wezel, M. C. [1995a] "Bankruptcy Prediction", *Proceedings of the INWGA*, Vaasa, Finland, January 1995, pp.285-300.
- Back, B., Oosterom, G., Sere, K. and Wezel, M. C. [1995b] "The Comparative Study of Neural Networks in Bankruptcy Prediction", Presented at the 18th Annual Congress of the European Accounting Association, Venice, Italy, April 1995.
- Bajgier, S. M. and Hill, A. V. [1982]. "An Experimental Comparison of Statistical and Linear Programming Approaches to the Discriminant Problem", *Decision Sciences*, 13, 1982, pp.604-618.
- Barnes, P. [1982] "Methodological Implications of Non-Normality Distributed Financial Ratios", *Journal of Business and Finance*, June 1984, pp.171-198.
- Barnes, P. [1987] "The Analysis and Use of Financial Ratios: A Review Article", *Journal of Business Finance and Accounting*, Winter 1987, pp.449-461.
- Barnett, V. D. and Lewis, T. [1978] "Outliers in Statistical Data", John Wiley & Sons, 1978.
- Barron, A. R. [1992] "Universal Approximation Bounds Superpositions of A Sigmoid Function", *IEEE Transaction on Information Theory* 38, 1992.
- Bates, T. [1973] "An Econometric Analysis of Lending to Black Business", *Review of Economics and Statistics*, August 1973.
- Baum, E. B. and Wilczek, F., [1988] "Supervised Learning of Probability Distributions by Neural Networks", In *Neural Information Processing Systems* (December, 1987), ed. D. Z. Anderson, pp.52-61.
- Beaver, W. H. [1966] "Financial Ratios as Predictors of Failure "Empirical Research in Accounting : Selected Studies", 1966, Supplement to *Journal of Accounting Research*, pp.71-111.
- Beaver, W. H. [1968] "Alternative Accounting Measures as Predictors of Failure" *The Accounting Review*, January 1968, pp.113-122.
- Beaver, W. H. Kettler, P. and Scholes, M. [1970] "The Association between Market Determined and Accounting Determined Risk Measures", *The Accounting Review*, October 1970, pp.654-682.
- Becker, S. and Le Cun, Y. [1989] "Improving the Convergence of Backpropagation Learning with Second Order Methods", In *Proceeding of the 1988 Connectionist Models Summer School* (Pittsburgh 1988), eds. D. Touretzky, G. Hinton and T. Sejnowski, pp.29-37. San. Mateo: Morgan Kaufmann.
- Belkaoui, A. [1980] "Industrial Bond Rating: A New Look", *Financial Management*, Autumn, 1980.

- Bell, T. B., Ribar, G. S. and Verchio, J. R. [1990] "Neural Nets vs. Logistic Regression: A Comparison of Each Model's Ability to Predict Commercial Bank Failures", Working Paper Presented to 1990 Deloitte & Touche/University of Kansas Auditing Symposium, May 1990.
- Berry, B. and Trigueiros, D. "Applying Neural Networks to the Extraction of Knowledge from Accounting Reports: A Classification Study", Trippi, R. & Turban, E. ed., *Neural Networks in Finance and Investing*. Probus, Chicago, pp.103-123.
- Bharadwaj, S., Flower, J. and Kolawa, A. [1992] "Issues in Dynamic Parallelization", *AI Expert*, Vol. 7, No. 2, February 1992, pp.27-33.
- Bharath, R. and Drosen, J. [1994] "Neural Network Computing", Windcrest/McGraw-Hill, 1994.
- Bird, R. G. and McHugh, A. J. [1977] "Financial Ratios – An Empirical Study", *Journal of Business Finance and Accounting*, Spring 1977, pp.29-45.
- Blum, M. [1974] "Failing Company Discriminant Analysis", *Journal of Accounting Research*, Spring 1974, pp.1-25.
- Blum, E. K. [1989] "Approximation of Boolean Functions by Sigmodial Networks: Part I: XOR and Other Two-Variable Functions", *Neural Computation*, 1, pp.532-540.
- Bougen, P. D. and Drury, J. C. [1980] "U.K. Statistical Distribution of Financial Ratios", *Journal of Business Finance and Accounting*, Spring 1980, pp.39-47.
- Brady, M., Raghavan, R. and Slawny, J. [1989] "Backpropagation Fails to Separate Where Perceptions Succeed", *IEEE Transactions on Circuits and Systems*, 36, pp.65-674.
- Bryson, A. E. and Ho, Y. C. [1969] "Applied Optimal Control", New York: Blaisdell, 1969.
- Buijink, W. and Jegers, M. [1986] "Cross-sectional Distributional Properties of financial Ratios in Belgian Manufacturing Industries: Aggregation Effects and Persistence Over Time", *Journal of Business Finance and Accounting*, Autumn 1986.
- Carleton, W. T. and Lerner, E. M. [1969] "Statistical Credit Scoring of Municipal Bonds", *Journal of Money, Credit, and Banking*, November 1969.
- Carpenter, G. A. and Grossberg, S. [1987] "ART 2: Self-Organization of Stable Category Recognition Codes for Analog Input Pattern", *Applied Optics*, 26, pp.4919-4930.
- Casey, C. J. and Bartczak, N. J. [1984] "Cash Flow- It's not the Bottom Line", *Harvard Business Review*, July-August 1984, pp.61-66.
- Casey, C. and Bartczak, N. [1985] "Using Operating Cash Flow Data to Predict Financial Distress: Some Extension", *Journal of Accounting Research*, Spring 1985, Vol.23, No.1. pp.384-401.
- Cater, J. P. [1987] "Successfully Using Peak Learning Rates of 10 (and Greater) in Back-Propagation Networks with the Heuristic Learning Algorithm", In *IEEE First International Conference on Neural Networks (San Diego 1987)*, eds. M. Caudill and C. Butler, Vol. II, pp.645-652, New York: IEEE.

- Caudill, M. [1991] "Evolutionary Neural Network", *AI Expert: The Magazine of Artificial Intelligence in Practice*, Vol.3, No.6. pp.53-59.
- Caudill, M. [1990] "Neural Network Primer", Miller Freeman Publications, San Francisco, California, 1990.
- Chatfield, C. [1993] "Neural Networks: Forecasting Breakthrough or Passing Fad?", *Internationally Journal of Forecasting*, 9, pp.1-3.
- Chen, K. H. and Shimerda, T. A. [1981] "An Empirical Analysis of Useful Financial Ratios", *Financial Management*, Spring 1981, pp.51-60.
- Chesser, D. [1974] "Predicting Loan Noncompliance", *The Journal of Commercial Bank Lending*, August 1974, pp.28-38.
- Chinganda, E. F. and Subrahmaniam, K. [1979] "Robustness of the Linear Discriminant Function to Nonnormality: Johnson's System", *Journal of Statistical Planning and Inference* 3, 1979, pp.167-179.
- Chua, W. [1986] "Radical Developments in Accounting Thought", *The Accounting Review*, Vol.61, No.4, pp.601-632.
- Coats, P. K. [1988] "Why Expert System Fail", *Financial Management*, August 1988, pp.77-86.
- Coats, P. K. and Fant, L. F. [1993] "Recognising Financial Distress Patterns Using A Neural Network Tool", *Financial Management*, November 1993, pp.142-155.
- Cochran, W. G. [1964] "On the Performance of the Linear Discriminant Function", *Technometrics*, 6:2, May 1964, pp.179-190.
- Cochran, W. G. [1963] "Sampling Techniques", John Wiley & Sons, 1963.
- Collins, R. and Green, R. [1982] "Statistical Methods for Bankruptcy Prediction", *Journal of Economics and Business* 34, 1982, pp.349-354.
- Cooley, W. W. and Lohnes, P. R. [1962] "Multivariate Procedures for the Behavioral Sciences", John Wiley and Sons: New York, 1962.
- Cooley, W. W. and Lohnes, P. R. [1971] "Multivariate Data Analysis", John Wiley and Sons: New York, 1971.
- Crawley, D. R. [1979] "Logistic Discrimination as An Alternative to Fisher's Linear Discriminant Function", *N. Z. Statis.* 14, pp.21-25.
- Cybenko, G. [1989] "Approximation by Superpositions of A Sigmodal Function,", *Mathematics of Control, Signals and Systems*, 1989, 2(4), pp.303-314.
- Dambolena, I. G. and Khoury, J. [1980] "Ratio Stability and Corporate Failure", *The Journal of Finance*, pp.1018-1026.
- Dambolena, I. G. and Shulman, J. M. [1988] "A Primary Rule for Detecting Bankruptcy: Watch the Cash", *Financial Analysts Journal*, September-October 1988, pp.74-78.



- Deakin, E. B. [1972] "A Discriminant Analysis of Predictions of Business Failure", *Journal of Accounting Research*, Spring, 1972, pp.167-179.
- Deakin, E. B. [1976] "Distributions of Financial Accounting Ratios: Some Empirical Evidence", *The Accounting Review*, January 1976, pp. 90-96.
- Denton, J. W., Hung, M. S. and Osyk, B. A. [1990] "A Neural Network Approach to the Classification Problem", *Expert System with Application*, Vol.1, No.4, 1990, pp.417-424.
- Diamond, H. S. Jr. [1976] "Pattern Recognition and the Detection of Corporate Failure", Unpublished Ph.D. Dissertation, New York University. 1976.
- Dopuch, N., Holthausen, R. W. and Leftwich, R. W. [1987] "Predicting Audit Qualification with Financial and Market Variables", *The Accounting Review*, July 1987, pp.4312-454.
- Dorsey, R. E., Johnson, J. D. and Mayer, W.J. [1994] "A Genetic Algorithm for the Training of Feedforward Neural Networks", *Advances in Artificial Intelligence in Economics, Finance and Management*, Vol.1, Edited by Andrew Whinston and John D. Johnson, JAI Press pp.93-111.
- Dunn, O. J. and Clark, V. A. [1987] "Applied Statistics: Analysis of Variance and Regression", New York: Wiley, 1987.
- Dutta, S. and Shekhar, S. [1988] "A Non-Conservative Application of Neural Networks", *ICNN, International Conference on Neural Networks*, San Diego, CA, July 1988, pp.24-27.
- Dwyer, M. D. [1992] "A Comparison of Statistical Techniques and Artificial Neural Network Models in Corporate Bankruptcy Prediction", Unpublished Ph.D Dissertation, The University of Wisconsin-Madison, 1992.
- Ederington, L. [1985] "Classification Models and Bond Rating", *The Financial Review*, 1985, November, 20.
- Edmister, R. O. [1972] "An Empirical Test of Financial Ratio Analysis for Small Business Failure Prediction", *Journal of Financial and Quantitative Analysis*, March 1972, pp.1477-1493.
- Efron, B. [1975] "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis", *Journal of the American Statistical Association* 70, pp.892-898.
- Eisenbeis, R. A., [1977] "Pitfall in the Application of Discriminant Analysis in Business and Economics", *The Journal of Finance*, June 1977, pp.875-900.
- Erxleben, k., Baetge, J., Feidicker, M., Koch, H., Krause, C. and Mertens, P. [1992] "Klassifikation Von Unternehmen. Ein Vergleich Vonneuronalen Netzen Und Diskriminanzanalyse"  $\square$  w", *Zeitschrift fur Betriebswirtschaft*, 11 1992, pp.1237-1262.
- Erxleben, K. and Koch, H. [1991] "Fruherkennung Von unternehmenskrisen-ein Vergleich Von Neuronalen Netzen and Diskriminanzanalyse", Working Paper, Universitat Erlangen-Nurnberg, Abteilung Wirtschaftsinformatik.
- Ezzamel, M., Mar-Molinero, C. and Beecher, A. [1987] "On the Distributional Properties of Financial Ratios", *Journal of Business Finance and Accounting*, Winter 1987, pp.463-482.

- Fahlman, S. E. [1989] "Fast-Learning Variations on Back-Propagation: An Empirical Study", in Proceeding of the 1988 Connectionist Models Summer School (Pittsburgh 1988), eds. D. Touretzky, G. Hinton and T. Sejnowski, pp.38-51. San. Mateo: Morgan Kaufmann.
- Fahlman, S. E. and Lebiere, C. [1990] "The Cascade-Correlation Learning Architecture", Technical Report: CMU-CS-90-100, Carnegie Mellen University, February 1990.
- Financial Accounting Standards Board [1987] "Statement of Cash Flows", Statement of Financial Accounting Standards No.95. Stamford, Connecticut: FASB, 1987.
- Fisher, R. A. [1936] "The Use of Multiple Measurements in Taxonomic Problem", Annals of Eugenics, 7, 1936, pp.179-188.
- Fitzpatrick, P. [1932] "A Comparison of the Ratios of Successful Industrial Enterprises with Those of Failed Companies", The Accountants Publishing Company.
- Fogel, D. B. [1991] "An Information Criterion for Optimal Neural Network Selection", IEEE Transactions on Neural Networks, Vol. 2, No.5, September 1991, pp.480-497.
- Foster, G. [1986] "Financial Statement Analysis", 2nd ed. Prentice-Hall International Editions, 1986.
- Frecka, T. J. and Hopwood, W. S. [1983] "The Effects of Outliers on the Cross-Sectional Distributional Properties of Financial Ratios", The Accounting Review, Vol. LVIII, No. 1, January 1983.
- Freed, N. and Glover, F. [1986] "Resolving Certain Difficulties and Improving the Classification Power of LP Discriminant Analysis Formulations", Decision Sciences, 17:4, 1986, pp.589-595.
- Frydman, H., Altman, E. I. and Kao, D. [1985] "Introducing Recursive Partitioning for Financial Classification: The Case of Financial Distress", Journal of Finance, 40, 1, 1985, pp.269-291.
- Fukunaga, K. and Hayes, R. [1989] "Estimation on Classifier Performance", IEEE Transactions on Pattern Analysis and Machine Intelligence, 1989, 11(10), pp.1087-1101.
- Funahashi, K. [1989] "On the Approximate Realization Continuous Mappings by Neural Networks", Neural Network, 2, 1989, pp.83-192.
- Gallinari, P., Thiria, S. and Fogelman, S. F. [1988] "Multilayer Perceptron and Data Analysis", IEEE International Conference on Neural Networks, Vol. 1, 1988, pp.391-399.
- Geman, S., Beienenstocks E. and Doursat, R. [1992] "Neural Networks and The Bias/ Variance Dilemma", Neural Computation, 4, 1992, pp.1-58.
- Gentry, J. A., Newbold, P. and Whitford, D. T. [1985a] "Classifying Bankruptcy Firms with Funds Flow Components", Journal of Accounting Research, Spring 1985, pp.146-160.
- Gentry, J. A., Newbold, P. and Whitford, D. T. [1985b] "Predicting Bankruptcy: If Cash Flow's not the Bottom Line, What is ?", Financial Analysts Journal, September-October 1985, pp.47-56.
- Gessner, G., Kamakura, W. A., Malhotra, N. K. and Zmijewski, M. E. [1988] "Estimating Models with Binary Dependent Variables: Some Theoretical and Empirical Observations", Journal of Business Research 16(1), 1988, pp.49-65.

- Gilbert, E. S. [1968] "On Discrimination Using Qualitative Variables", *American Statistical Association Journal*, 63:324, December 1968, pp.1399-1412.
- Gilbert, E. S. [1969] "The Effect of Unequal Variance-Covariance Matrices on Fisher's Linear Discriminant Function", *Biometrics*, 25:3, September 1969, pp.505-515.
- Goldberg, D. [1989] "Genetic Algorithms in Search, Optimisation, and Machine Learning", Addison-Wesley, Reading, MA, 1989.
- Gombola, M. J., Haskins, M. E., Ketz J. E. and William D. D. [1987] "Cash Flow in Bankruptcy Prediction", *Financial Management*, Winter 1987, pp.55-65.
- Gordon, T. [1974] "Hazards in the Use of the Logistic Function with Special Reference to Data from Prospective Cardiovascular Studies", *Journal of Chronic Diseases*, 27, 1974, pp.97-102.
- Gori, M. and Tesi, A. [1992] "On the Local Minimum in Backpropagation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(1), pp.76-86.
- Gorman, R. P. and Sejnowski, T. P. [1988] "Analysis of Hidden Units in A Layered Network Trained to Classify Sonar Targets", *Neural Network*, 1, pp75-89.
- Gorr, W. L., Nagin, D. and Szczypula, J. [1994] "Comparative Study of Artificial Neural Network and Statistical Models for Predicting Student Grade Point Averages", *International Journal of Forecasting*, 10 1994, pp.17-34.
- Greene, D. P. [1987] "Automated Knowledge Acquisition: Overcoming the Expert System Bottleneck", *Proceedings of the 8th International Conference on Information System*, 1987, pp.107-117.
- Grier, P. and Katz, S. [1976] "The Differential Effects of *Bond Rating Changes among Industrial and Public Utility Bonds* by Maturity", *Journal of Business*, 49, 1976, pp.226-239.
- Grudnitski, G. and Do, A. Q. [1993] "Important Factors in Neural Networks' Forecasts of Futures Prices", *Refenes, Apostolos-Paul ed., Neural Networks in the Capital Markets* (John Wiley & Sons), 1995.
- Halperin, M., Blackwelder, W. C. and Verter, J. I. [1971] "Estimation of the Multivariate Logistic Risk Function: A Comparison of the Discriminant and Maximum likelihood Approaches", *J. Chron. Dis.* 24, pp.125-158.
- Hamer, M. M. [1983] "Failure Prediction: Sensitivity of Classification Accuracy to Alternative Statistical Methods and Variable Sets", *Journal of Accountig and Public Policy*, 1983, pp.289-307.
- Hand, D. J. [1981] "Discrimination and Classification", New York: John Wiley & Sons.
- Hansen, J. V. and Messier, W. F. [1986] "A Preliminary Investigation of EDP-Expert", *Auditing: A Journal of Practice and Theory*, Fall 1986, pp.109-123.
- Hanson, S. J. and Pratt, L., [1989] "A Comparison of Different Bias for Minimal Neural Network Construction with Back-Propagation", *Advances in Neural Information Processing System*, ed. by D. S. Touretzky, Vol.1, San.Mateo, CA: Morgan Kanfmann, pp.177-185.

- Hawley, D. D., John, J. D. and Raina, D. [1990] "Artificial Neural Systems: A New Tool for Financial Decision-Making", *Financial Analysts Journal*, Vol.46, No.6, November /December 1990, pp.63-72.
- Hayes-Roth, Frederick, Waterman, D. A. and Lenat, D. B. [1983] "Building Expert System", Reading, M.A.: Addison-Wesley, 1983.
- Hayter, A. J. [1984] "A Proof of the Conjecture That the Tukey-Kramer Method is Conservative", *The Annals of Statistics*, 12, 1984, pp. 61-75.
- Hecht-Nielsen, R. [1989] "Theory of the Backpropagation Neural Network", *Proceedings of the International Joint Conference on Neural Networks*, Vol.1, Washington, D.C., 1989, pp. 593-605.
- Helfert, E. A. [1982] "Techniques in Financial Analysis", 5th ed., Homewood, Ill. : Richard D. Irwin, 1982.
- Hertz, J., Krogh, A. and Palmer, R. G. [1991] "Introduction to the Theory of Neural Computation", Addison-Wesley Publishing Company, 1991.
- Hill, T., Marquez, L., O'Connor, M. and Remus, W. [1994] "Artificial Neural Network Models for Forecasting and Decision Making", *International Journal of Forecasting* 10 1994, pp.5-15.
- Hirose, Y., Yamashita, K. and Hijiya, S. [1991] "Back-Propagation Algorithm Which Varies the Number of Hidden Units", *Neural Networks*, Vol.4, No. 1, 1991, pp.61-65.
- Hopwood, W., McKewon, J. and Mutchler, J. [1988] "The Sensitivity of Financial Distress Prediction Models to Departures from Normality", *Contemporary Accounting Research*, Vol. 5 No. 1, Fall 1988, pp.284-298.
- Hopfield, J. J. [1987] "Learning Algorithms and Probability Distributions in Feed-Forward and Feed-Back Networks", *Proceedings of the National Academy of Science, USA*, 84, pp.8429-8433.
- Hornik, K., Stinchcombe, M. and White, H. [1990] "Universal Approximation of An Unknown Mapping and Its Derivatives Using Multilayer Feedforward Networks", *Neural Networks*, Vol.3, No.5, 1990, pp.551-560.
- Hornik, K. [1991] "Approximation Capabilities of Multilayer Feedforward Networks", *Neural Networks*, Vol.4, No.2, 1991, pp.251-257.
- Horrigan, J. O. [1965] "Some Empirical Bases of Financial Ratio Analysis", *The Accounting Review*, July 1965, pp.558-568.
- Horrigan, J. O. [1966] "The Determination of Long Term Credit Sharing with Financial Ratios", *Journal of Accounting Research*, 1966, 2, pp.44-62.
- Horrigan, J. O. [1968] "A Short History of Financial Ratio Analysis", *The Accounting Review*, April 1968, pp.284-294.
- Horton Jr, J. J. [1970] "Statistical Classification of Municipal Bonds", *Journal of Bank Research*, 1, Autumn 1970.

- Hosmer, D. W. and Lemeshow, S. [1989] "Applied Logistic Regression", New York: John Wiley & Sons.
- Hosmer, T. A., Hosmer, D. W. and Fisher, L. [1983a] "A Comparison of the Maximum Likelihood and Discriminant Function Estimators of the Coefficients of the Logistic Regression Model for Mixed Continuous and Discrete Variables", *Communications in Statistics — Computation and Simulation*, 12, pp.23-43.
- Hosmer, T. A., Hosmer, D. W. and Fisher, L. [1983b] "A Comparison of Three Methods of Estimating the Logistic Regression Coefficients", *Communications in Statistics -Computation and Simulation*, 12, pp.577-593.
- Hsieh, S. J. [1993] "A Note on the Optimal Cut-off Point in Bankruptcy Prediction Models", *Journal of Business Finance and Accounting*, Vol. 20, No.3, April 1993. pp.457-464.
- Huberty, C. J. [1984] "Issues in the Use and Interpretation of Discriminant Analysis", *Psychological Bulletin*, Vol.95, 1984, pp.156-171.
- Hudson, J. [1986] "Analysis of Company Liquidation", *Applied Economics*, 1986, pp.22-25.
- Izan, H. Y. [1984] "Corporate Distress in Australia", *Journal of Banking and Finance*, 1984, Vol. 8, pp.303-320.
- Jacobs, R. A. [1988] "Increased Rates of Convergence through Learning Rate Adaptation. *Neural Networks* 1, pp.295-307.
- Joachimsthaler, E. A. and Stam, A. [1988] "Four Approaches to the Classification Problem in Discriminant Analysis: An Experimental Study", *Decision Sciences*, 19, 1988, pp.322-333.
- Jobson, J. D. [1992] "Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods", Springer-Verlag New York, Inc.
- Johnson, C. G. [1970] "Ratio Analysis and the Prediction of Firm Failure", *The Journal of Finance*, December 1970, pp.1166-1172.
- Johnson, M. [1987] "Multivariate Statistical Simulation", John Wiley & Sons, 1987.
- Johnson, P. E. [1983] "What Kind of Expert Should A System be?", *The Journal of Medicine and Philosophy*, Vol.8, No.1, February 1983, pp.77-97.
- Jones, F. L. [1987] "Current Techniques in Bankruptcy Prediction", *Journal of Accounting Literature* Vol 6., 1987, pp.131-164.
- Jordan, M. I. [1986] "An Introduction to Linear Algebra in Parallel Distributed Processing", In Rumelhart, D. E. and McClelland, J. L., *Parallel Distributed Processing Volume I: Foundations*, MIT Press. Cambridge, Massachusetts, 1986.
- Joshi, P. C. [1972] "Efficient Estimation of the Mean of An Exponential Distribution When An Outlier is Present", *Technometrics*, February 1972, pp.137-144.
- Joy, O. M. and Tollefson, J. O. [1975] "On the Financial Applications of Discriminant Analysis", *Journal of Financial and Quantitative Analysis*, December 1975, pp.723-739.

- Judge, G., Hill, R. Griffiths, W., Lutkepohl, and Lee, T. [1982] "Introduction to the Theory and Practice of Econometrics", John Wiley & Sons, New York, 1982.
- Jutten, C., Guerin, A. and Nguyen Thi, H. L. [1991] "Adaptive Optimization of Neural Algorithms", In LNCS 540, A. Prieto (Ed.), Artificial Neural Networks, pp.54-61. New York: Springer-Verlag.
- Kale, B. K. [1974] "Detection of Outliers", Technique Report No.63, Department of Statistics, University of Winnipeg, Canada.
- Kale, B. K. [1975] "Trimmed Means and the Method of Maximum Likelihood When Spurious Observations are Present", In R. P. Gupta (ed.), Applied Statistics, North Holland, 1975.
- Kaplan, R. and Urwitz, U. [1979] "Statistical Model of Bond Ratings : A Methodological Inquiry", Journal of Business, April 1979, pp.231-261.
- Karels, G. V. and Prakash, A. [1987] "Multivariate Normality and Forecasting of Business Bankruptcy", Journal of Business Finance and Accounting, Winter 1987, pp.573-593.
- Keasey, K. and McGuinness, P. [1990] "The Failure of UK Industrial Firms for the Period 1976-1984, Logistic Analysis and Entropy Measure", Journal of Business Finance and Accounting, 1990, Vol.17, No.1.
- Keasey, K. and Waston, R. [1991] "Financial Distress Prediction Models: A Review of Their Useness", British Journal of Management, 1991, Vol.2, pp.89-102.
- Keppel, G. [1982] "Design and Analysis: A Researcher's Handbook", Second Edition. Englewood Cliff: Prentice-Hall, inc., 1982.
- Keyes, J. [1992] "Living in Parallel", AI Expert, Vol, 7, No. 2, February 1992, pp42-47.
- Kim, J. and Muller, C. W. [1978] "Introduction to Factor Analysis – What It is and How to Do It", Quantitative Applications in the Social Sciences Series. Beverly Hills, Ca.: Sage Publications, 1978.
- Kim, J. W., Weistroffer, H. R. and Redmond, R. T., [1993] "Expert System for Bond Rating: A Comparative Analysis of Statistical, Rule-Based and Neural Network Systems", Expert Systems, August 1993, Vol, 10, No.3, pp.167-171.
- Kirk, R. E. [1968] "Experimental Design: Procedures for the Behaviour Sciences", Brooks/Cole Publishing Co., 1968.
- Kleinbaum, D. G., Kupper, L. L. and Muller, K. E. [1988] "Applied Regression Analysis and Other Multivariate Methods", 2nd ed., Boston, Ma.: PWS-Kent, 1988.
- Klimisaukas, C. C., Guiver, J. and Pelton, G. [1989] "Neural Computing", Pittsburgh: Neural Ware Inc.
- Knight, K. [1990] "Conectionist, Ideas, and Algorithms", Communications of the ACM, Vol.33, No.11, pp.59-74.

- Koh, H. C. [1987] "Prediction of Going-Concern Status: A Probit Model for the Auditors", Unpublished Dissertation, Virginia Polytechnic Institute and State University, 1987.
- Koh, H. C. [1992] "The Sensitivity of Optimal Cut-off Points to Misclassification Costs of Type I and Type II Errors of the Going Concern Prediction Context", *Journal of Business Finance and Accounting*, 19(2) January, 1992, pp.187-197.
- Kohonen, T. [1989] "Self-Organization and Associative Memory", 3rd ed, Springer-Verlag, Berlin.
- Kohonen, T. [1990] "The Self-Organizing Map", *Proc. of the IEEE*, 78, 9, pp.1464-1480.
- Kolen, J. F. and Pollack, J. B. [1990] "Back Propagation is Sensitive to Initial Conditions", TR 90-JK-BPSIC, Lab for AI Research, Ohio State University, 1990.
- Kolmogorov, A. N. [1968] "On the Representation of Continuous Function of Many Variables by Superposition of Continuous Functions of One Variable and Addition", *American Mathematical Society Translation*, Vol.28. pp.55-59.
- Korobow, D. J. and Stuhr, D. P. [1975] "Toward Early Warning of Changes in Banks' Financial Condition: A Progress Report", *Federal Reserve of New York Monthly Review*, 1975, pp.157-165.
- Krogh, A., Thorbergsson, G. I. and Hertz, J. A. [1990] "A Cost Function for Internal Representations", *Advances in Neural Information Processing Systems II (Denver 1989)*, ed. D. S. Touretzky, pp.733-740. San. Mateo: Morgan Kaufmann.
- Kullback, S. [1959] "Information Theory and Statistics", New York: John Wiley & Sons.
- Lachenbruch, P. A. [1967] "An Almost Unbiased Method of Obtaining Confidence Intervals for The Probability of Misclassification Models in Discriminant Analysis", *Biometrics*, December 1967, pp.639-645.
- Lachenbruch, P. A. and Mickey, M. R. [1968] "Estimation of Error Rates in Discriminant Analysis", *Technometrics*, February 1968, pp.1-11.
- Lachenbruch, P. A., Sneeringer, C. and Revo, L. T. [1973] "Robustness of the Linear and Quadratic Discriminant Function to Certain Types of Non-Normality", *Communications in Statistics*, 1 : 1, 1973, pp.39-56.
- Laitinen, E. K. [1991] "Financial Ratios and Different Failure Processes", *Journal of Business Finance and Accounting*, September 1991.
- Lang, K. J. and Witbrock, M. J. [1988] "Learning to Tell Two Spiral Apart", In *Proceedings of the 1988 Connectionist Models Summer School*, D. S. Touretzky and T. J. Sejnowski, eds. San Mateo, CA: Morgan Kaufmann Publishers, 1988.
- Largay, J. A. III and Stickney, C. P. [1980] "Cash Flows, Ratio Analysis and the W.T. Grant Company Bankruptcy", *Financial Analysts Journal*, July-August 1980, pp.51-54.
- Lawson, G. H. [1985] "The Management of Performance on A Cash Flow Basis, A Reply to Mr. Eggington", *Accounting and Business Research*, Spring 1985, pp.85-104.

- Lee, C. [1985] "Stochastic Properties of Cross-Sectional Financial Data", *Journal of Accounting Research*, Spring 1985, pp.213-227.
- Lee, Y., Oh, S. H. and Kim, M. W. [1991] "The Effect of Initial Weight on Performance Saturation in Back-Propagation Learning", In *Pro.1991 Int. Joint Conf. Neural Networks*, Seattle, WA, July 8-12, 1991, pp.I-765-770.
- Lee, C. K. and Ord, J. K. [1990] "Discriminant Analysis Using Least Absolute Deviation", *Decision Science*, 21, 1990, pp.86-96.
- Lev, B. [1974] "Financial Statement Analysis: A New Approach", Prentice Hall.
- Levitan, A. S. and Knoblett, J. A. [1985] "Indicators of Exceptions to the Going Concern Assumption", *Auditing: A Journal of Practice & Theory*, Fall 1985, pp.26-39.
- Levitt, M. E. [1995] "Machine Learning for Foreign Exchange Trading", Refenes, A. N. ed., *Neural Networks in the Capital Markets*, pp.301- 307.
- Lewis, T., and Fieller, N. R. J. [1979] "A Recursive Algorithm for Null Distributions for Outliers: I. Gamma Samples", *Technometrics*, August 1979, pp.187-210.
- Lin, L. H. [1993] "An Examination of Stability of Forecasting in Failure Prediction Models", Unpublished Ph.D Thesis. University of Warwick , Warwick Business School, 1993.
- Lippmann, R. P. [1987] "An Introduction to Computing with Neural Nets", *IEEE ASSP Magazine*, April 1987, pp4-22.
- Lippmann, R. P. and Beckman, P. [1989] "Adaptive Neural Net Processing for Signal Detection in Non-Gaussian Noise", *Advances in Neural Information Processing Systems*, edited by Touretzky, David., Vol.1, San Mateo, CA:Morgan Kaufmann, 1989, pp.124-132.
- Maddala, G. S. [1983] "Limited-Dependent and Qualitative Variables in Economics", Cambridge University Press, 1983.
- Maddala, G. S. [1988] "Introduction to Econometrics", Macmillan Publishing Comapny.
- Manski, C. F. and Lerman, S. R. [1977] "The Estimation of Choice Probabilities from Choice Based Samples", *Econometrica* 45, November 1977, pp.1977-1988.
- Manski, C. F. and McFadden, D. [1981] "Structural Analysis of Discrete Data with Econometric Applications", Cambridge, MA: MIT Press, 1983.
- Mardia, K. V. [1970] "Measures of Multivariate Skewness and Kurtosis with Application", *Biometrika* 57, pp.519-520.
- Marks, S. and Dunn, O. J. [1974] "Discriminant Functions When Covariance Matrices Are Unequal", *Journal of the American Statistical Association*, 9:346, June 1974, pp.55-559.
- Martin, D. [1977] "Early Warning of Bank Failure — A Logit Regression Approach", *Journal of Banking and Finance*, 1977, pp.249-276.



- Mason, R. J. and Harris, F. C. [1979] "Predicting Company Failure in the Construction Industry", Proceedings of the Institution of civil Engineers 66, Part 1, May, pp.301-307.
- Masters, T. [1993] "Practical Neural Network Recipes in C++", New York: Academic Press.
- McCullagh, P. and Nelder, J. A. [1989] "Generalised Linear — Models", 2 Edition, New York: Chaman Hallm.
- McDonald, B. and Morris, M. [1984] "The Statistical Validity of the Ratio Method in Financial Analysis: An Empirical Examination", Journal of Business Finance & Accounting, Spring 1984.
- McDonald, B. and Morris, M. [1985] "The Functional Specification of Financial Ratios: An Empirical Examination", Accounting and Business Research, Summer 1985, pp.223-228.
- McFadden, D. [1976] "A Comment on Discriminant Analysis 'versus' Logit Analysis", Annals of Economic and Social Measurements, 5, 1976, pp.511-523.
- McLachlan, G. J. [1992] "Discriminant Analysis and Statistical Pattern Recognition", New York: John Wiley & Sons.
- McLachlan, G. J. and Basford, K. E. [1988] "Mixture Model", New York: Marcel Deckker. Inc.
- McLeay, S. [1986a] "Student's t and the Distribution of Financial Ratios", Journal of Business Finance and Accounting, 13(2), Summer 1986, pp.209-222.
- McLeay, S. [1986b] "The Ratio of Means, the Mean of Ratios and Other Benchmarks", Finance, Journal of the French Finance Society, 1986.
- McLeay, S. and Fieldsend, S. [1987] "Sector and Size Effects in Ratio Analysis – An Indirect Test of Ratio Proportionality", Accounting and Business Research, Spring 1987, pp.228-245.
- Mecimore, C. D. [1968] "Some Empirical Distributions of Financial Ratio", Management Accounting USA, September 1968, pp.13-16.
- Meehl, P. E., and Rosen, A. [1955] "Antecedent Probability and the Efficiency of Psychometric Signs, Patterns or Cutting Score", Psychological Bulletin, Vol. 52, No.3, 1955, pp.194-216.
- Mendenhall, W. and Scheafer, R. L. [1973] "Mathematical Statistics with Applications", Wadsworth Publishing Co., 1973.
- Mensah, Y. M. [1983] "The Differential Bankruptcy Predictive Ability of Specific Price Level Adjustments: Some Empirical Evidence", The Accounting Review, 1983, pp.228-245.
- Mensah, Y. M. [1984] "An Examination of the Stationarity of Multivariate Bankruptcy Prediction Models: A Methodological Study", Journal of Accounting Research, Vol. 22 1984, pp.380-395.
- Merwin, C. [1942] "Financing Ratios and the Probabilistic Prediction of Bankruptcy", Journal of Accounting of Illinois, Bureau of Business Research.
- Miller, A. J. [1990] "Subset Selection in Regression", London: Chapman and Hall, 1990.
- Minsky, M. and Papert, S. [1969] "Perceptrons", Cambridges, MA: MIT Press, 1969.

- Monroe, R. J. and Simkowitz, M. A. [1971], "Investment Characteristics of Conglomerate Targets: A Discriminant Analysis", *Southern Journal of Business*, November 1971.
- Moore II, D. H. [1973] "Evaluation of Five Discrimination Procedures for Binary Variables", *Journal of American Statistical Association* 68, 1973, pp.339-404.
- Morrison, D. F. [1969] "On the Interpretation of Discriminant Analysis", *Journal of Marketing Research*, May 1969, pp.156-163.
- Mosteller, F. and Wallace, D. L. [1963] "Inferences in the Authorship Problem", *Journal of the American Statistical Association* 1963, Vol.58, No.302, pp.275-309.
- Mount, K. S. and Kale, B. K. [1973] "On Selecting A Spurious Observation", *Canadian Mathematical Bulletin*, March 1973, pp.75-78.
- Mozer, M. C. and Smolensky, P. [1989] "Skeletonization: A Technique for Trimming the Fast From A Network via Relevance Assessment", *Advances in Neural Information Processing System I* (Denver 1988), San. Mateo: Morgan Kaufmann.
- Mutchler, J. F. [1985] "A Multivariate Analysis of the Auditor's Going-Concern Opinion Decision", *Auditing: A Journal of Practice & Theory*, Spring 1984, pp.668-682.
- Myers, R. H. [1986] "Classical and Modern Regression with Applications", Boston: Duxbury Press.
- Nelder, J. A. and Wedderburn, R. W. M. [1972] "Generalised Linear Model", *Journal of the Royal Statistical Society, Series A*, 135, pp.761-768.
- Neter, J. "Discussion of Financial Ratios as Predictors of Failure", *Empirical Research in Accounting: Selected Studies*, 1966. Supplement to Vol. 4, *Journal of Accounting Research*, pp.112-118.
- Neural Computing [1993] - A Technology Handbook for Professional II/PLUS and NeuralWorks Explorer. Pittsburgh, United States of America: NeuralWare, 1993, pp.218-219.
- Neuralware Professional II/PLUS Software. Pittsburgh, United States of America: NeuralWare.
- Nolen, J. N. [1992] "Parallel Processing for Problem Solving", *AI Expert*, Vol.7, No.2, February 1992, pp.35-40.
- Nygard, K., Ficek, R. and Sharda, R. [1992] "Genetic Algorithm", *OR/MS Today*, August 1992, pp.28-34.
- O'Connor, M. C. [1973] "On the Usefulness of Financial Ratios to Investors in Common Stock", *The Accounting Review*, April 1973, pp.339-352.
- Odom, D .M. and Sharda, R. [1989] "A Neural Network Model for Bankruptcy Prediction" In *Proceedings of the International Joint Conference on Neural Network*, 1990.
- O'Hara, T. F., Hosmer, D. W., Lemeshow, S. and Hartz, S. C. [1982] "A Comparison of Discriminant Function and Maximum Likelihood Estimates of Logistic Coefficients for Categorical-Scaled Data", *Journal of Statistical Computation and Simulation*, 14, pp.169-178.

- Ohlson, J. A. [1980] "Financial Ratios and the Probabilistic Prediction of Bankruptcy", *Journal of Accounting Research*, Spring 1980, pp.109-131.
- Park, Y. R. [1993] "A Study of Neural Network Construction and Its Application to Statistical Forecasting", Unpublished Ph.D. Dissertation, Syracuse University, 1993.
- Parker, D. B. [1985] "Learning Logic", Technical Report TR-47, Center for Computational Research in Economics and Management Science, Massachusetts Institute of Technology, Cambridge, MA.
- Parker, D. B. [1987] "Optimal Algorithms for Adaptive Networks: Second Order Back Propagation, Second Order Direct Propagation, and Second Order Herbian Learning", In *IEEE First International Conference on Neural Networks* (San Diego 1987), eds. M. Caudill and C. Butler, Vol.II, pp.593-600.
- Peavy, J. [1984] "Forecasting Industrial Bond Rating Changes: A Multivariate Approach", *Review of Business and Economic Research*, 19, 1984, pp.46-56.
- Peavy, J. W. and Scott, J. A. [1986] "The AT&T Diverstiture: Effects of Rating Changes on Bond Returns", *Journal of Economics and Business*, 38,1986, pp.255-270.
- Peel, M. J. [1987] "Timeliness of Private Company Accounts and Predicting Corporate Failure: A Logistic Analysis", *Investment Analysts*, 1987, No.83, pp23-27.
- Pinches, G. and Singleton, J. [1978] "The Adjustment of Stock Prices to Bond Rating Changes", *Journal of Finance*, 33 1978, pp.29-44.
- Pinches, G. E. and Mingo, K. A. [1973] "A Multivariate Analysis of Industrial Bond Ratings", *Journal of Finance*, March 1973, pp.1-18.
- Pinches, G. E. Mingo, K. A. and Caruthers, J. K. [1973] "The Stability of Financial Patterns in Industrial Organizations", *Journal of Finance*, May 1973, pp.389-396.
- Pinches, G. E., Eubank, A. A., Mingo, K. A. and Caruthers, J. K. [1975] "The Hierarchical Classification of Financial Ratios", *Journal of Business Research*, Vol.3, No.4, October 1975, pp.295-310.
- Pinches, G. E. [1980] "Factors Influencing Classification Results from Multiple Discriminant Analysis", *Journal of Business Research*, December 1980, pp.429-456.
- Pindyck, R. and Rubinfeld, D. [1981] "Econometric Model & Economic Forecasts", 2 Edition. McGraw-Hill.
- Platt, H. D. and Platt, M. B. [1990], "Development of A Class of Stable Predictive Variables: The Case of Bankruptcy Prediction", *Journal of Business Finance & Accounting*, 1990.
- Plaut, D. S., Nowlan, S. and Hinton, G. [1986] "Experiments on Learning by Back Propagation", Technical Report CMU-CS-86-126, Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

- Poddig, T. [1992] "Kunstliche Interigenz und Entscheidungstheorie", Lehrstuhl für Finanzwirtschaft und Banken, Feiburg, Wiesbaden, pp.604-618.
- Press, J. and Wilson, S. [1978] "Choosing between Logistic Regression and Discriminant Analysis", Journal of the American Statistical Association. December 1978, pp.699-705.
- Rahimian, E., Singh, S., Thammachote, T. and Virmani, R. [1993] "Bankruptcy Prediction by Neural Network", Trippi, R. R. & Turban, E. ed., Neural Networks in Finance and Investing, Probus, Chicago, pp.159-171.
- Ramser, J. and Foster, L. [1931] "A Demonstration of Ratio Analysis", Bulletin No. 40, Urbana, Ill. University of Illinois, Bureau of Business Research.
- Refenes, A. N. [1994] "Comments on Neural Networks: Forecasting Breakthrough or Passing Fad by C. Chatfield", International Journal of Forecasting 10, 1994, pp.43-46.
- Refenes, A. N., Zapranis, A. and Francis, G. [1994] "Stock Performance Modeling Using Neural Networks: A Comparative Study with Regression Models", Neural Networks, Vol.7, No.2, 1994, pp.375-388.
- Reilly, D. L., Cooper, L. N., Elbaum, C. [1982] "A Neural Model for Category Learning", Biological Cybernetics, 45, pp.35-41.
- Riply, B. D. [1993] "Statistical Aspects of Neural Networks" in Barndorff-Nielsen, O.E., Jensen, J. L. and Kendall, W. S., eds., Networks and Chaos: Statistical and Probabilistic Aspects, London: Chapman & Hall.
- Rose, P. S., Andrews, T. and Giroux, G. A. [1982] "Predicting Business Failure: A Macro-Economic Perspective", Journal of Accounting, Auditing & Finance, pp.20-31.
- Rosenblatt, F. [1959] "Two Theorems of Statistical Separability in the Perceptron", Mechanisation of Thought Processes: Proceedings of a Symposium held at the National Physical Laboratory, Vol. 1, London: HM Stationary Office, 1959, pp.421-456.
- Rosenblatt, F. [1962] "Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanism", Washington, D.C., Spartan Books, 1962.
- Rulon, P. J., Tiedeman, D. V., Tatsuoka, M. M. and Langmuir, C. R. [1967] "Multivariate Statistics for Personal Classification", John Wiley and Sons, Inc., New York, 1967.
- Rumelhart, D. E., Hinton, G. E. and Williams, R. J. [1986] "Learning Internal Representations by Error Propagation", In Parallel Distributed Processing: Explorations in the Microstructure of Cognition, edited by Rumelhart, D. E., McClelland J. L. and PDP Research Group, Vol.1, Cambridge: MIT Press. 1986, pp.319-362.
- Salchenberger, L. M., Cinar, E. M. and Lash, N. A. [1992] "Neural Networks: A New Tool for Predicting Failure", Decision Science, Vol.24, No.4, July/August 1992, pp.899-916.
- Samdani, G. [1990] "Neural Nets: They Learn from Examples", Chemical Engineering, Vol. 97(8), August 1990, pp.37-45.

- Santomero, A .M. and Vinso, J. D. [1977] "Estimating the Probability of Failure for Commercial Banks and the Banking System", *Journal of Banking and Finance*, September 1977, pp.182-205.
- Sarle, W. S. [1994] "Neural Network and Statistical Models", *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, April 1994.
- Scott, E. [1978] "On the Financial Application of Discriminant Analysis Comment", *Journal of Financial and Quantitative Analysis*, March 1978, pp.201-205.
- Scott, J. Jr. [1976] "A Theory of Optimal Capital Structure", *Bell Journal of Economics*, 7, 1, pp.33-54.
- Scott, J. Jr. [1977] "Bankruptcy, Secured Debt, and Optimal Capital Structure", *Journal of Finance*, 32, 1, pp.1-14.
- Scott, J. Jr. [1981] "The Probability of Bankruptcy: A Comparison of Empirical Predictions and Theoretical Models", *Journal of Banking and Finance*. 5, pp.317-344.
- Sen, T., Oliver, R., and Sen, N. [1992] "Predicting Corporate Mergers Using Backpropagation Neural Networks, A Comparative Study with Logistic Models", *Virginia Tech. R. B. Pamplin College of Business, Department of Accounting*.
- Serrano, C., Martin, B. and Galizo, J. L. [1993] "Artificial Neural Network in Financial Statement Analysis: Ratios versus Accounting Data", *The 16th Annual Congress of the European Accounting Association*, Turku, Finland, April 1993, pp.28-30.
- Shadmehr, R. and D'argenio, D. Z. [1990] "A Comparison of A Neural Network Based Estimator and Two Statistical Estimators in A Sparse and Noisy Data Environment", *Proceedings of the International Joint Conference on Neural Networks*, Vol.1, Washington, D.C., 1990, pp.289-292.
- Sharda, R. and Wilson, R. [1993] "Performance Comparison Issues in Neural Network Experiments for Classification Problems", Nunamaker E. and Sprague, R. ed., *Organisational System and Technology: Proceedings of the 26 Hawaii Int. Conf. on System Sciences (IV)*, (IEEE Computer Society Press, 1993), pp.649-657.
- Sheth, J. N. [1979] "How to Get the Most out of Multivariate Methods", In *Multivariate Data Analysis*, Hair, Anderson, Tatham, and Grablovsky (eds.), Tulsa, OK, Petroleum Publishing Company, 1979, Ch.1.
- Sietsma, J. and Dow, R. J. [1989] "Neural Net Pruning — Why and How" *Proceedings of the International Joint Conference on Neural Networks*, Vol. 1, San Diego, CA: 1988, pp.325-333.
- Sietsma, J. and Dow, R. J. [1991] "Creating Artificial Neural Networks That Generalise", *Neural Networks*, Vol. 4, No.1, 1991, pp.67-79.
- Singleton, J. and Surkan, A. [1990] "Neural Networks for Bond Rating Improved by Multiple Hidden Layers", in *Proceedings of the IEEE International Conference on Neural Networks*, 1990 Vol. II, pp.163-168.
- Singleton, J. and Surkan, A. [1993] "Bond Rating with Neural Networks", Refenes, A. N. ed., *Neural Networks in the Capital Markets*, pp.301- 307.

- Sinha, S. K. [1972] "Reliability Estimation in Life Testing in the Presence of An Outlier Observation", *Operation Research*, July-August, 1972, pp.888-894.
- Sinha, S. K. [1973a] "Distributions of Order Statistics and Estimation of Mean Life When an Outlier May be Present", *Canadian Journal of Statistics*, Vol. 1, No.1, 1973, pp.235-243.
- Sinha, S. K. [1973b] "Lifetesting and Reliability Estimation for Non-Homogeneous Data — A Bayesian Approach", *Communications in Statistics*, Vol.2. No.3, 1973, pp.235-243.
- Sinha, S. K. [1973c] "Estimation of Parameters of Two Parameter Exponential Distribution When an Outlier May be Present", *Utilitas Mathematica*, Vol. 3, 1973, pp.75-82.
- Smith, C. A .B. [1947] "Some Examples of Discrimination ", *Annals of Eugenics*, 13, 1947, pp.272-282.
- Solla, S. A., Levin., E. and Fleisher, M. [1988] "Accelerated Learning in Layered Neural Networks", *Complex System 2*, pp. 625-639.
- Sontag, E. D. and Sussmann, H. J. [1989] "Backpropagation Can Give Rise to Spurious Local Minimum even for Networks without Hidden Layers", *Complex Systems*, 3, pp.91-106.
- Specht, D. F. [1990] "Probabilistic Neural Networks", *neural networks*, 3, pp109-118.
- Stam, A. and Jones, D. G. [1990] "Classification Performance of Mathematical Programming Techniques in Discriminant Analysis: Results for Small and Medium Sample Size", *Managerial and Decision Economics*, 11, 1990, pp.243-253.
- Stam, A. and Joachimsthaler, E. A. [1990] "A Comparison of A Robust Mixed-Integer Approach to Existing Methods for Establishing Classification Rules for the Discriminant Problem", *European Journal of Operational Research* , 46, 1990, pp.113-122.
- Steele, A. [1995] "Going Concern Qualifications and Bankruptcy Prediction", *Working Paper*, University of Warwick, 1995.
- Stevens, D. L. [1973] "Financial Characteristics of Merged Firms: A Multivariate Analysis", *Journal of Financial and Quantitative Analysis*, March 1973, pp.149-158.
- Storey, D., Keasey, K., Watson, R. and Wynarczyk, P. [1987] "The Performance of Small Firms", *Croon-Helm*, Bromley.
- Sudarsanam, P. S. and Taffler, R. J. [1985] "Industrial Classification in U.K. Capital Markets: A Test of Economics Homogeneity", *Applied Economics*, 1985.
- Surkan, A. J., and Singleton, J. C. [1989] "Neural Network for Bond Rating Improved by Multiple Hidden Layers", *IJCNN-89*, Vol.2, pp.157-162.
- Tabachnick, B. G. and Fidell, L. S. [1989] "Using Multivariate Statistics", *Second Edition*. New York: Horper and Row, 1989.
- Taffler, R. J. [1982] "Forecasting Company Failure in the UK Using Discriminant Analysis and Financial Ratio Data", *Journal of Royal Statistics Society*, [1982], 145, Part 3, pp.342-358.

- Tam, K. Y. and Kiang, M. Y. [1990] "Predicting Bank Failure: A Neural Network Approach", *Applied Artificial Intelligence*, 4, 1990, pp.265-282.
- Tam, K. Y. and Kiang, M. Y. [1992] "Managerial Application of Neural Networks: The Case of Bank Failure Prediction", *Management Science*, Vol.38, No.7, July 1992, pp.926-947.
- Tatsuoka, M. M. [1971] "Multivariate Analysis: Techniques for Educational and Psychological Research", New York: John Wiley and Sons, Inc., 1971.
- Thunhurst, C. [1985] "The Analysis of Small Area Statistics and the Planning for Health", *The Statistics*, Vol.34, 1985.
- Tollefson, J. O. and Joy, O. M. [1978] "Some Clarifying Comments on Discriminant Analysis", *Journal of Financial and Quantitative Analysis*, March 1978, pp.197-200.
- Trigueiros, D. and Taffler, R. [1995] "Neural Networks and Empirical Research in Accounting", Working Paper at the City University Business School, London, 1995.
- Utans, J. and Moody, J. [1991] "Selecting Neural Network Architecture via the Prediction Risk: Application to Corporate Bond Rating Prediction", *Proc. First International Conference on Artificial Intelligence Applications on Wall Street*, New York, U.S.A. 1991, pp.35-41.
- Veal, J. R. and Kale, B. K. [1972] "Tests of Hypotheses for Expected Life in the Presence of A Spurious Observation", *Utilitas Mathematica*, November 1972, pp.9-23.
- Von Lehmen, V., Paek, E. G., Liao, P. E., Marrakchi, A., and Patel, J. S. [1988] "Factors Influencing Learning by Backpropagation", *IEEE International Conference on Neural Network*, Vol.1 1988, pp.335-341.
- Wahl, P. W. and Kronwal, R. A. [1977] "Discriminant Functions When Covariance are Unequal and Sample Sizes Moderate", *Biometrics*, Vol.33, September 1977, pp.479-484.
- Wasserman, P. D. [1989] "Neural Computing: Theory and Practice", Van Nostrand Reinhold, 1989.
- Wasserman, P. D. and Schwartz, T. [1987] "Neural Networks Part 1", *IEEE Expert*, Vol.2, Winter 1987, pp.10-13.
- Watson, C. [1990] "Multivariate Discriminant Properties, Outlier and Transformation of Financial Ratios", *The Accounting Review*, July 1990, pp.682-695.
- Watts, R. L. and Zimmerman, J. L. [1986] "Positive Accounting Theory", Prentice-Hall, 1986.
- Weems, Charles C., Brown, C., Webb, J., Poggio, T. and Kender, J. R. [1991] "Parallel Processing in the DARPA Strategic Computing Version Program", *IEEE Expert*, Vol. 6, No. 5 October 1991, pp.23-28.
- Weisberg, S. [1985] "Applied Linear Regression" New York : John Wiley & Sons.
- Weiss, S. M. and Kulikowski, C. A. [1991] "Computer Systems That Learn", Morgan Kaufmann Publishers", San Mateo, California, 1991.

- Werbos, P. J. [1974] "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences", Ph.D Dissertation, Harvard University, 1974.
- West, R. [1970] "An Alternative Approach to Predicting Corporate Bond Ratings", *Journal of Accounting Research*, 7, 1970, pp.118-127.
- White, H. [1992] "Artificial Neural Networks: Approximation and Learning Theory", Oxford UK: Blackwell.
- White, H. [1989] "Some Asymptotic Results for Learning in Single Hidden Layer Feedforward Network Models", *Journal of the American Statistical Association*, 84, 1989, pp.1003-1013.
- Wilcox, J. [1971a] "A Gambler's Ruin Prediction of Business Failure Using Accounting Data", *Sloan Management Review*, Spring 1971, 12, 3, pp.1-11.
- Wilcox, J. [1971b] "A Simple Theory of Financial Ratios as Predictors of Failure", *Journal of Accounting Research*, Autumn 1971, pp.389-395.
- Wilcox, J. [1973] "A Prediction of Business Failure Using Accounting Data", *Empirical Research in Accounting, Selected Studies, Supplement to Vol.11, Journal of Accounting Research*, 1973, pp.163-179.
- Wilcox, J. [1976] "The Gamblers Ruin Approach to Business Risk", *Sloan Management Review*, Fall 1976, pp.33-46.
- Wilensky, G. D. and Manukian, N. [1992] "The Projection Neural Network", *IJCNN*, 1992, Vol.II, pp.358-367.
- Wilson, R. L. and Sharda, R. [1994] "Bankruptcy Prediction Using Neural Networks", *Decision Support Systems* 11(1994), pp545-557.
- Winakor, A. and Smith, R. [1935] "Changes in the Financial structure of Unsuccessful Industrial corporations", *Bulletin No. 51*, 1935. University of Illinois, Bureau of Business Research: Urbana, Illinois.
- Wittner, B. S. and Denker, J. S. [1988] "Strategies for teaching Layered Networks Classification Tasks", In *Neural Information Processing System* (Denver 1987), ed. D. Z. Anderson, pp.850-859. New York: American Institute of Physics.
- Wong, F., Wang, P., Goh, T. and Quek, B. [1992] "Fuzzy Neural Systems for Stock Selection", *Financial Analysts Journal*, Jan/Feb 1992.
- Yoon, Y. and Swales, G. [1993] "Predicting Stock Price Performance: A Neural Network Approach", *Neurovest Journal*, 1(1), pp.14-15.
- Yoon, Y., Swales, G. and Margavio, T. [1993] "A Comparison of Discriminant Analysis Versus Artificial Neural Networks", *Journal of the Operational Research Society*, 1993, Vol.44, No.1, pp.51-60.
- Zavgren, C. V. [1983] "The Prediction of Corporate Failure: The State of the Art", *Journal of Accounting Literature*, Vol.2, 1983.

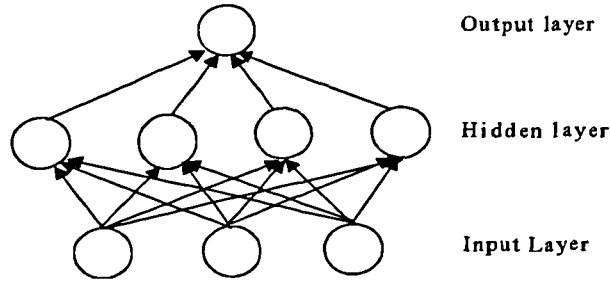


- Zavgren, C. V. [1985] "Assessing the Vulnerability to Failure of American Industrial Firms: A Logistic Analysis", *Journal of Business Finance & Accounting* , Spring 1985, 12(1).
- Zavgren, C., Michael, V., Dugan, T. and Reeve, J. M. [1988] "The Association between Probability of Bankruptcy and Market Responses - A Test of Market Anticipation", *Journal of Business Finance & Accounting*, Spring 1988, 15(1).
- Zhezhel, Y. N. [1968] "The Efficiency of A Linear Discriminant Function for Arbitrary Distributions", *Engineering Cybernetics*. No.6, 1968, pp.107-111.
- Zmijewski, M. [1983] "An Empirical Comparison of the Extant Financial Distress Prediction Models", Unpublished Working Paper, University of Chicago.
- Zmijewski, M. [1984] "Methodological Issues Related to the Estimation of Financial Distress Prediction Models", *Journal of Accounting Research*, Supplement, 1984, Vol.22, pp59-82.

## APPENDIX I

### The Mathematical Derivation of BP Algorithm

Consider a multi-layer neural network illustrated by the following figure



A Multi-Layer Neural Network

Using notation

$A_j^{[n]}$  = the output of processing element  $j$  in layer  $n$

$Z_j^{[n]}$  = the combined weighted summation of processing element  $j$  in layer  $n$

$W_{ij}^{[n]}$  = the weight connected processing element  $i$  in layer  $n-1$  to the element  $j$  in layer  $n$

$\theta_j^{[n]}$  = the threshold of processing element  $j$  in layer  $n$

$\delta_j^{[n]}$  = the bias of processing element  $j$  in layer  $n$  when receiving  $i$  value from the preceding layer

$T_j$  = the desired value of element  $j$  in output layer when receiving value from preceding layer

$\Delta W_{ij}^{[n]}$  = the change in weight connected processing element  $i$  in layer  $n-1$  to element  $j$  in layer  $n$

$\Delta \theta_j^{[n]}$  = the change in threshold of processing element  $j$  in layer  $n$

$\alpha$  = the momentum.  $0 < \alpha < 1$

$\eta$  = learning rate

$f(x)$  = transfer function

Suppose a weighted input for the processing element between the hidden layer and the output layer is

$$Z_j^{[n]} = W_{ij}^{[n]} A_j^{[n-1]} - \theta_j^{[n]}$$

An output value will be produced through transferring  $Z_j$  by transfer function. As was mentioned before, in the backpropagation neural network sigmoid is the most frequently used transfer function.

Thus the network output value  $A_j$  of processing element  $j$  in layer  $n$  will be

$$A_j^{[n]} = f(Z_j^{[n]}) = (1 + e^{-Z})^{-1}$$

Since the purpose of a supervised learning network is to minimise the error between computed output value and desired output value, we set a global error function (cost function)  $E$  associated with it which is a differentiable function of all the connection weights in the network in order to measure the quality of learning. Our usual global error measure or cost function is

$$\text{Min } E = \frac{1}{2} \sum (T_j - A_j)^2$$

The gradient steepest descent method is applied to minimise the above equation because of its continuous differentiable attribute. It implies that the network will adjust the value of weight when it learns each time from the input of the training sample. The adjustment is proportional to the difference between the desired output and the actual system output. That is, the change in the weight will be commensurate with the error function sensitivity to the weights and can be expressed in the following form by using chain rule

$$\Delta W_{ij}^{[n]} = -\eta \frac{\partial E}{\partial W_{ij}^{[n]}} = -\eta \left( \frac{\partial E}{\partial Z_j^{[n]}} \right) \left( \frac{\partial Z_j^{[n]}}{\partial W_{ij}^{[n]}} \right) = -\eta \underbrace{\left( \frac{\partial E}{\partial A_j^{[n]}} \right)}_{(1)} \underbrace{\left( \frac{\partial A_j^{[n]}}{\partial Z_j^{[n]}} \right)}_{(2)} \underbrace{\left( \frac{\partial Z_j^{[n]}}{\partial W_{ij}^{[n]}} \right)}_{(3)}$$

The term (1) (2) and (3) is broken down into different components;

$$(2) \frac{\partial A_i^{[n]}}{\partial Z_j^{[n]}} = \frac{\partial}{\partial Z_j^{[n]}} f(Z_j) = f'(Z_j^{[n]})$$

$$(3) \frac{\partial Z_j^{[n]}}{\partial W_{ij}^{[n]}} = \frac{\partial}{\partial W_{ij}^{[n]}} f(Z_j^{[n]}) = \frac{\partial}{\partial W_{ij}^{[n]}} f\left(\sum_k W_{kj}^{[n]} A_k^{[n-1]} - \theta_j^{[n]}\right) = A_i^{[n-1]}$$

(1) The first term can be further divided into the following two situations

(a) the layer n is the output layer

$$\frac{\partial E}{\partial A_i^{[n]}} = \frac{\partial}{\partial A_i^{[n]}} \left[ \frac{1}{2} \sum_k (T_k - A_k^{[n]})^2 \right] = -(T_i - A_i^{[n]})$$

(b) the layer n is the hidden layer. From the chain rule, we obtain

$$\frac{\partial E}{\partial A_i^{[n]}} = \sum_k \left( \frac{\partial E}{\partial Z_k^{[n+1]}} \right) \left( \frac{\partial Z_k^{[n+1]}}{\partial A_i^{[n]}} \right)$$

we define

$$\frac{\partial E}{\partial Z_k^{[n+1]}} = -\delta_k^{[n+1]} \quad \text{and} \quad \frac{\partial Z_j^{[n+1]}}{\partial A_i^{[n]}} = \frac{\partial}{\partial A_i^{[n]}} (\sum W_{ik} A_i^{[n]} - \theta_k) = W_{ik}$$

thus, (1) becomes

$$\frac{\partial E}{\partial A_j^{[n]}} = -\sum_k \delta_k^{[n+1]} W_{jk} f(Z_j^{[n]})$$

In summary, the gradient descent rule gives

1. For the hidden-to-output connections, the gradient descent rule gives

$$\Delta W_{ij} = -\eta \frac{\partial E}{\partial W_{ij}} = -\eta (T_j - A_j^{[n]}) f'(Z_j^{[n]}) A_i^{[n-1]} = \eta \delta_j^{[n]} A_i^{[n-1]}$$

$$\text{where } \delta_j^{[n]} = -(T_j - A_j^{[n]}) f'(Z_j^{[n]})$$

2. For the input to hidden connections

$$\Delta W_{ij} = -\eta \frac{\partial E}{\partial W_{ij}} = -\eta [\sum_k \delta_k^{[n+1]} W_{jk}] f'(Z_j^{[n]}) A_i^{[n-1]} \eta \delta_j^{[n+1]} A_i^{[n-1]}$$

$$\text{where } \delta_j^{[n]} = [-\sum_k \delta_k^{[n+1]} W_{jk}] f'(Z_j^{[n]})$$

Therefore, whatever  $W_{ij}$  is located between the hidden layer and the output layer or between the hidden layer and hidden layer,  $\Delta W_{ij}$  can be written in the same form

$$\Delta W_{ij} = \frac{\partial E}{\partial W_{ij}} = \eta \delta_j^{[n]} A_i^{[n-1]}$$

where

$A_i^{[n-1]}$  = the output value from the preceding layer connected by  $W_{ij}$

$\delta_j^{[n]}$  = the difference value connected by  $W_{ij}$  to the higher layer

$$= \begin{cases} -(T_j - A_j^{[n]}) f'(Z_j^{[n]}) & \text{if in the output layer} \\ -\sum_k \delta_k^{[n+1]} W_{jk} f'(Z_j^{[n]}) & \text{otherwise} \end{cases}$$

## APPENDIX II

### Input Data for Example 1

x1	x2	G	x1	x2	G	x1	x2	G	x1	x2	G
4.69	26.63	1	6.92	24.29	1	3.57	23.46	0	7.45	17.92	0
5.46	25.82	1	6.91	24.30	1	7.18	18.31	0	6.50	19.28	0
3.72	27.65	1	7.28	23.90	1	7.81	17.41	0	8.88	15.88	0
4.03	27.33	1	6.77	24.45	1	7.69	17.58	0	7.87	17.32	0
5.02	26.28	1	5.95	25.31	1	7.25	18.21	0	4.73	21.80	0
5.07	26.23	1	4.03	27.32	1	7.97	17.18	0	2.05	25.63	0
5.65	25.62	1	6.55	24.68	1	4.22	22.53	0	6.92	18.67	0
7.63	23.54	1	5.46	25.82	1	8.97	15.74	0	5.35	20.91	0
5.01	26.29	1	4.98	26.33	1	9.19	15.43	0	5.34	20.93	0
4.18	27.17	1	6.51	24.72	1	7.26	18.19	0	5.81	20.26	0
6.05	25.20	1	5.93	25.33	1	3.72	23.25	0	5.55	20.63	0
5.25	26.04	1	7.90	23.25	1	3.46	23.62	0	6.21	19.69	0
3.51	27.88	1	6.97	24.23	1	4.71	21.83	0	3.59	23.43	0
5.28	26.01	1	5.04	26.26	1	5.94	20.07	0	8.45	16.48	0
4.90	26.41	1	5.87	25.39	1	6.44	19.37	0	6.85	18.77	0
5.84	25.43	1	7.43	23.75	1	7.43	17.95	0	5.40	20.85	0
7.44	23.74	1	5.97	25.29	1	10.17	14.03	0	8.53	16.37	0
8.18	22.95	1	6.26	24.98	1	5.93	20.08	0	7.48	17.88	0
6.19	25.05	1	6.30	24.94	1	7.34	18.07	0	2.96	24.33	0
4.47	26.87	1	6.38	24.85	1	6.31	19.55	0	3.85	23.06	0
5.83	25.44	1	4.61	26.71	1	5.66	20.48	0	4.19	22.58	0
4.48	26.86	1	7.88	23.27	1	5.34	20.93	0	9.00	15.70	0
4.58	26.75	1	4.93	26.38	1	7.65	17.63	0	7.90	17.27	0
2.80	28.62	1	4.61	26.71	1	3.66	23.33	0	4.77	21.74	0
6.43	24.81	1	4.54	26.79	1	7.71	17.54	0	6.49	19.29	0
5.84	25.42	1	3.77	27.60	1	8.65	16.21	0	4.97	21.47	0
9.37	21.71	1	6.36	24.88	1	6.17	19.74	0	4.99	21.43	0
6.62	24.60	1	5.84	25.42	1	6.81	18.83	0	8.45	16.49	0
8.47	22.66	1	4.90	26.41	1	5.62	20.53	0	5.82	20.25	0
6.15	25.10	1	5.22	26.08	1	5.25	21.06	0	5.85	20.20	0
6.03	25.22	1	4.86	26.45	1	9.22	15.39	0	4.44	22.22	0
5.44	25.84	1	5.65	25.62	1	10.42	13.67	0	4.03	22.81	0
3.75	27.62	1	5.42	25.87	1	7.99	17.14	0	4.93	21.52	0
5.79	25.48	1	6.87	24.34	1	10.27	13.88	0	11.17	12.61	0
5.91	25.35	1	5.14	26.15	1	3.03	24.24	0	5.31	20.97	0
8.65	22.47	1	4.10	27.26	1	5.00	21.42	0	6.00	19.98	0
3.51	27.88	1	7.52	23.65	1	6.70	18.99	0	7.95	17.21	0
7.01	24.19	1	7.20	23.99	1	7.13	18.37	0	6.18	19.73	0
6.68	24.53	1	5.44	25.85	1	5.84	20.22	0	5.07	21.32	0
5.68	25.59	1	6.57	24.66	1	8.04	17.08	0	5.86	20.19	0
6.51	24.72	1	5.47	25.81	1	2.10	25.55	0	5.17	21.17	0
5.73	25.53	1	6.73	24.48	1	5.55	20.63	0	9.83	14.52	0
5.42	25.86	1	6.09	25.15	1	5.51	20.70	0	6.04	19.94	0
5.03	26.28	1	7.14	24.06	1	6.79	18.87	0	5.53	20.67	0
6.53	24.69	1	4.80	26.52	1	8.35	16.63	0	4.17	22.61	0
5.81	25.45	1	5.62	25.65	1	4.10	22.70	0	6.86	18.76	0
5.79	25.47	1	4.73	26.59	1	6.98	18.58	0	8.52	16.39	0
6.16	25.08	1	5.11	26.19	1	7.22	18.25	0	7.53	17.80	0
4.52	26.81	1	6.98	24.22	1	8.36	16.62	0	4.53	22.08	0
6.93	24.28	1	7.34	23.85	1	9.66	14.76	0	5.77	20.31	0

G denotes group

## APPENDIX III

### Input Data for Example 2

x1	x2	x3	G	x1	x2	x3	G
5.6860	5.2240	24.5155	1	8.0420	4.8640	9.4755	0
6.5160	6.5710	21.9626	1	2.1090	5.0560	16.6758	0
5.7390	5.8800	23.7113	1	5.5550	4.0930	13.4516	0
5.4260	5.1690	24.9024	1	5.5100	5.4290	12.0049	0
5.0320	6.8170	23.5409	1	6.7900	4.6820	11.2453	0
6.5370	2.9400	26.0213	1	8.3540	4.6780	9.2948	0
5.8140	4.8470	24.7796	1	4.1050	6.1170	12.9871	0
5.7960	5.3980	24.1823	1	6.9890	4.3440	11.3768	0
6.1660	6.3860	22.6083	1	7.2230	5.3240	9.9818	0
4.5220	5.4830	25.6791	1	8.3650	5.3720	8.5002	0
6.9320	7.5050	20.3919	1	9.6650	5.1230	7.1554	0
4.6990	5.3190	25.6424	1	3.5750	6.1370	13.6271	0
5.4640	3.6000	26.6200	1	7.1800	3.7100	11.8513	0
3.7290	3.9400	28.4063	1	7.8130	4.8380	9.7910	0
4.0350	5.2320	26.5703	1	7.6910	5.8820	8.7690	0
5.0270	3.5990	27.1674	1	7.2500	6.3620	8.7802	0
5.0750	2.8300	27.9725	1	7.9720	4.4740	10.0018	0
5.6560	4.5680	25.2910	1	4.2290	5.3590	13.6849	0
7.6320	5.2790	22.0211	1	8.9760	5.2240	7.9030	0
5.0180	3.0490	27.7974	1	9.1970	4.0070	8.9959	0
4.1800	1.9380	30.0948	1	7.2650	4.1500	11.2500	0
6.0550	4.5740	24.7855	1	3.7200	4.7040	15.0580	0
5.2570	2.6960	27.8958	1	3.4640	4.8210	15.2464	0
3.5120	5.4960	26.9270	1	4.7180	4.5760	13.9545	0
5.2830	5.8130	24.3566	1	5.9460	5.2120	11.7040	0
4.9080	5.0230	25.7141	1	6.4400	5.6940	10.5443	0
5.8410	6.7330	22.6241	1	7.4300	5.0810	9.9964	0
7.4440	4.8550	22.7331	1	10.1730	6.2230	5.2829	0
8.1890	1.9500	25.0700	1	5.9390	3.3650	13.7906	0
6.1940	5.6680	23.3810	1	7.3460	5.1150	10.0631	0
4.4730	4.7530	26.5616	1	6.3120	5.4870	10.9371	0
5.8320	4.0910	25.6076	1	5.6620	5.3570	11.8959	0
4.4800	1.0960	30.6670	1	5.3460	4.1210	13.6814	0
4.5830	4.7830	26.3904	1	7.6580	5.1060	9.6832	0
2.8070	3.7150	29.8119	1	3.6670	3.9820	15.9365	0
6.4300	5.1500	23.6688	1	7.7160	5.1930	9.5129	0
5.8420	5.3600	24.1675	1	8.6510	5.7980	7.6635	0
9.3750	5.5360	19.5533	1	6.1760	5.8490	10.6999	0
6.6280	5.4020	23.1378	1	6.8160	4.7570	11.1284	0
8.4720	3.4070	23.0771	1	5.6290	4.2970	13.1296	0
6.1520	3.3410	26.0514	1	5.2560	4.7140	13.1268	0
6.0350	5.1780	24.1310	1	9.2230	4.9120	7.9452	0
5.4450	4.7430	25.3579	1	10.4250	4.9500	6.4000	0
3.7540	4.1720	28.1140	1	7.9990	5.7500	8.5325	0
5.7920	2.2540	27.7243	1	10.2790	6.2440	5.1267	0
5.9100	5.9700	23.3963	1	3.0320	5.9310	14.5376	0
8.6520	5.5780	20.4098	1	5.0030	4.5330	13.6466	0
3.5130	5.2200	27.2363	1	6.7050	3.6590	12.5024	0
7.0110	5.2140	22.8705	1	7.1350	6.2030	9.1029	0
6.6880	1.7870	27.1296	1	5.8410	5.7620	11.2165	0

# APPENDIX III

(continue)

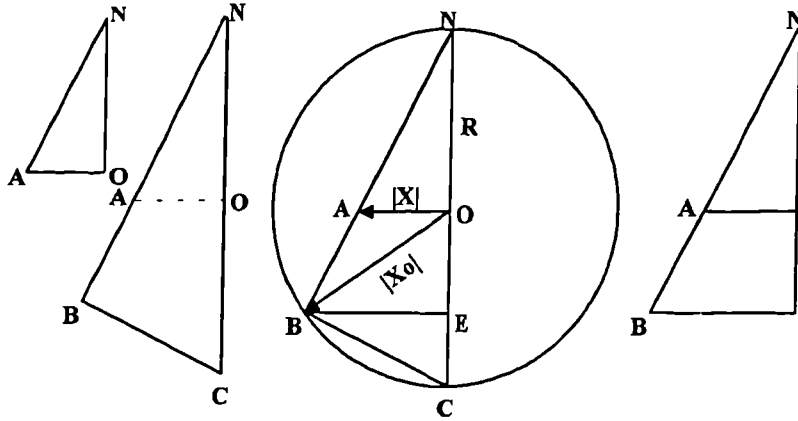
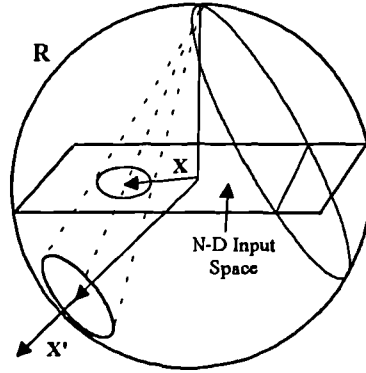
## Input Data for Example 2

x1	x2	x3	G	x1	x2	x3	G
6.5700	2.8210	26.1139	1	5.8660	4.3130	12.8154	0
5.4780	5.9760	23.9295	1	5.1750	6.3410	11.3976	0
6.7360	5.8120	22.5415	1	9.8350	6.5410	5.3476	0
6.0990	3.7730	25.6316	1	6.0410	6.0530	10.6391	0
7.1410	3.6050	24.5181	1	5.5310	4.8700	12.6075	0
4.8030	5.7150	25.0669	1	4.1720	5.5150	13.5806	0
5.6250	2.0160	28.2008	1	6.8670	4.6070	11.2334	0
4.7320	2.0710	29.2551	1	8.5230	4.8140	8.9305	0
5.1160	5.5160	24.8995	1	7.5380	5.5690	9.3124	0
6.9860	2.2680	26.2160	1	4.5390	5.4810	13.1601	0
7.3400	5.4730	22.1679	1	5.7780	3.9090	13.3799	0
6.9230	6.2000	21.8713	1	7.4500	4.9260	10.1458	0
6.9110	4.2150	24.1194	1	6.5000	5.6000	10.5750	0
7.2870	4.2650	23.5931	1	8.8810	4.4170	8.9296	0
6.7710	5.4630	22.8904	1	7.8740	5.4000	9.0825	0
5.9510	5.8190	23.5149	1	4.7390	4.1210	14.4401	0
4.0370	1.1420	31.1690	1	2.0540	5.5260	16.2158	0
6.5540	4.7560	23.9570	1	6.9250	5.6240	10.0168	0
5.4670	6.0950	23.8094	1	5.3590	3.9200	13.8913	0
4.9830	-.4370	31.7629	1	5.3460	5.9300	11.6463	0
6.5120	7.2400	21.2150	1	5.8160	5.3750	11.6831	0
5.9320	3.7580	25.8573	1	5.5530	5.1080	12.3123	0
7.9070	6.3800	20.4388	1	6.2150	6.0690	10.4036	0
6.9780	4.4890	23.7274	1	3.5950	5.4440	14.3818	0
5.0440	5.1720	25.3765	1	8.4590	3.8370	10.1096	0
5.8770	4.2180	25.4085	1	6.8560	5.3760	10.3820	0
7.4340	5.4250	22.1044	1	5.4050	4.3130	13.3916	0
5.9730	4.6780	24.7710	1	8.5380	5.1760	8.5045	0
6.2600	5.7280	23.2310	1	7.4810	5.6160	9.3308	0
6.3040	5.8150	23.0781	1	2.9690	5.5170	15.0821	0
6.3840	6.3070	22.4246	1	3.8570	4.0890	15.5786	0
4.6190	1.9630	29.5179	1	4.1920	4.5570	14.6334	0
7.8880	4.6300	22.4313	1	9.0070	5.5120	7.5403	0
4.9350	3.7820	27.0765	1	7.9090	6.4390	7.8699	0
4.6180	2.8290	28.5449	1	4.7760	4.5380	13.9248	0
4.5480	2.2580	29.2748	1	6.4930	6.0140	10.1180	0
3.7710	4.3900	27.8475	1	4.9710	5.6570	12.4221	0
6.3640	4.2520	24.7615	1	4.9990	4.6020	13.5740	0
5.8470	6.2010	23.2151	1	8.4540	6.2920	7.3540	0
4.9060	6.2120	24.3790	1	5.8250	4.7470	12.3784	0
5.2210	6.9540	23.1505	1	5.8550	5.1270	11.9134	0
4.8680	4.7480	26.0735	1	4.4410	5.3870	13.3884	0
5.6530	2.9250	27.1431	1	4.0330	4.8680	14.4823	0
5.4220	2.9720	27.3790	1	4.9300	6.0930	11.9829	0
6.8750	7.5240	20.4418	1	11.1730	6.9040	3.2668	0
5.1490	5.0690	25.3611	1	5.3180	4.1820	13.6478	0
4.1020	3.9340	27.9468	1	6.0090	5.7950	10.9694	0
7.5290	5.9910	21.3489	1	7.9530	5.6920	8.6552	0
7.2060	3.5220	24.5303	1	6.1870	5.5930	10.9741	0
5.4400	2.9970	27.3284	1	5.0740	4.6480	13.4285	0

G denotes group

## APPENDIX IV

### The Mathematical Derivation of an Alternative Projection



$$\triangle NAO \sim \triangle NCB$$

$$\frac{NA}{NC} = \frac{NO}{NB}$$

$$NB = \frac{NC \times NO}{NA} = \frac{2R^2}{\sqrt{R^2 + |X|^2}}$$

$$\triangle NAO \sim \triangle NBE$$

$$\frac{NA}{NB} = \frac{AO}{BE} = \frac{NO}{NE}$$

$$BE = \frac{NB \times AO}{NA} = \frac{2R^2 |X|}{R^2 + |X|^2} = R \frac{2R|X|}{R^2 + |X|^2}$$

$$NE = \frac{NB \times NO}{NA} = \frac{2R^2}{R^2 + |X|^2}$$

$$OE = NE - ND = R \frac{R^2 - |X|^2}{R^2 + |X|^2}$$

Therefore we obtain

$$\mathbf{X}' = R \left[ \frac{2R|X|}{R^2 + |X|^2}, \frac{R^2 - |X|^2}{R^2 + |X|^2} \right]$$



## APPENDIX V

### The Z sores for each of 264 companies and Descriptive Statistics by Group Using Multivariate Discriminant Analysis(MDA)

OBS	GROUP*	Z score	OBS	GROUP*	Z score	OBS	GROUP*	Z score
1	0	1.7777	49	0	10.3594	97	1	-2.1096
2	0	4.2268	50	0	6.3965	98	1	-2.1195
3	0	6.6339	51	0	6.5729	99	1	-2.4460
4	0	0.5753	52	0	3.2412	100	1	-3.3956
5	0	1.5694	53	0	1.4813	101	1	-4.5167
6	0	3.5762	54	0	10.0528	102	1	-2.2871
7	0	4.4076	55	0	5.2322	103	1	-4.1004
8	0	10.5408	56	0	2.8680	104	1	-4.0595
9	0	3.9464	57	0	7.1636	105	1	-0.9071
10	0	9.0478	58	0	2.6959	106	1	-3.6090
11	0	4.3147	59	0	11.5342	107	1	-1.8067
12	0	5.1436	60	0	10.1448	108	1	-0.7510
13	0	1.6876	61	0	6.2690	109	1	-1.3914
14	0	0.7598	62	0	0.1029	110	1	-3.8175
15	0	-0.0635	63	0	0.1046	111	1	-4.8978
16	0	2.6637	64	0	5.1676	112	1	-5.9282
17	0	1.1897	65	0	3.8243	113	1	-6.2153
18	0	-1.8386	66	0	1.6897	114	1	-3.7643
19	0	0.1034	67	0	1.6478	115	1	-2.2921
20	0	-0.3237	68	0	3.9882	116	1	-5.3823
21	0	0.2793	69	0	1.6266	117	1	-4.5530
22	0	0.7184	70	0	-0.9119	118	1	-5.0963
23	0	2.6635	71	0	2.3881	119	1	-1.1908
24	0	1.6560	72	0	5.1752	120	1	-3.1496
25	0	2.7940	73	0	5.3573	121	1	-4.8092
26	0	-1.3826	74	0	-0.0652	122	1	-3.0910
27	0	5.1969	75	0	4.6822	123	1	-1.5663
28	0	4.0835	76	0	1.2687	124	1	-3.5210
29	0	3.9582	77	1	17.6668	125	1	-4.9769
30	0	3.9463	78	1	1.2520	126	1	-5.8818
31	0	0.3132	79	1	3.1697	127	1	-5.9333
32	0	4.9245	80	1	0.9190	128	1	-0.5974
33	0	2.3319	81	1	1.9681	129	1	-2.3767
34	0	-0.2777	82	1	1.7544	130	1	-5.6268
35	0	-2.4320	83	1	4.4421	131	1	-4.7253
36	0	1.7248	84	1	-1.4558	132	1	-2.8068
37	0	4.1836	85	1	5.2551	133	1	-3.2500
38	0	1.6739	86	1	0.0673	134	1	-5.3801
39	0	6.0256	87	1	6.9385	135	1	-5.7659
40	0	6.1596	88	1	0.0515	136	1	-2.9206
41	0	15.7122	89	1	-4.4459	137	1	-2.2995
42	0	3.6558	90	1	-1.4345	138	1	-3.3732
43	0	1.6753	91	1	-4.1449	139	1	-6.3360
44	0	4.0177	92	1	-2.9541	140	1	-5.4933
45	0	4.9021	93	1	-2.2078	141	1	-4.0914
46	0	5.5227	94	1	-6.1412	142	1	-4.4283
47	0	6.2498	95	1	-0.5085	143	1	-2.3763
48	0	6.5882	96	1	-5.4494	144	1	-8.5485

# APPENDIX V

(continue)

OBS	GROUP*	Z score	OBS	GROUP*	Z score	OBS	GROUP*	Z score
145	1	-2.6382	185	1	-0.9744	225	1	-2.7500
146	1	-4.3066	186	1	-2.5198	226	1	-2.4001
147	1	-0.8220	187	1	-4.6504	227	1	-3.5082
148	1	-4.2346	188	1	-4.2208	228	1	-5.1946
149	1	-2.1488	189	1	-4.6877	229	1	-3.6733
150	1	-2.5510	190	1	-3.8267	230	1	-1.5370
151	1	-5.3384	191	1	-1.7948	231	1	-0.0069
152	1	-5.3096	192	1	-4.3585	232	1	-3.4505
153	1	-2.9079	193	1	-2.8207	233	1	-7.5177
154	1	-2.2135	194	1	-2.6365	234	1	-3.1169
155	1	-3.7054	195	1	-1.4730	235	1	-4.3400
156	1	-5.2809	196	1	-2.8576	236	1	-3.5554
157	1	-5.3119	197	1	-3.4560	237	1	-2.7720
158	1	-3.4682	198	1	-3.6942	238	1	-4.6010
159	1	-2.1271	199	1	-4.9015	239	1	-3.4049
160	1	-5.6207	200	1	-2.7295	240	1	-5.3340
161	1	-5.7142	201	1	-1.2647	241	1	-3.7566
162	1	-5.1440	202	1	-8.1465	242	1	-3.4090
163	1	-3.2450	203	1	-0.6919	243	1	-3.0317
164	1	-5.8795	204	1	-11.222	244	1	0.7270
165	1	-6.2585	205	1	-4.2922	245	1	-1.7720
166	1	-5.2933	206	1	-5.5796	246	1	-8.3771
167	1	-2.8527	207	1	-3.8558	247	1	-2.9401
168	1	-2.5473	208	1	-1.7740	248	1	-3.5941
169	1	-4.4196	209	1	-5.8906	249	1	-5.5759
170	1	-4.7203	210	1	-4.3720	250	1	-7.5468
171	1	2.5925	211	1	-7.8157	251	1	-4.8508
172	1	-2.8715	212	1	-3.2729	252	1	-2.1398
173	1	-5.3730	213	1	1.5647	253	1	-6.2906
174	1	-1.9037	214	1	4.1124	254	1	-6.4164
175	1	-3.1417	215	1	0.1896	255	1	-5.4317
176	1	-0.3728	216	1	-2.7986	256	1	-0.7152
177	1	-3.4749	217	1	-5.2808	257	1	-4.6315
178	1	-5.3735	218	1	-2.3182	258	1	-5.1306
179	1	-4.0030	219	1	-2.5540	259	1	-5.3565
180	1	-5.9255	220	1	-9.0226	260	1	-2.4630
181	1	-5.7721	221	1	-5.2942	261	1	-1.1418
182	1	-3.3524	222	1	-0.6862	262	1	0.1704
183	1	0.5286	223	1	-2.3378	263	1	-4.2072
184	1	-1.3876	224	1	-2.8821	264	1	-4.6018

\* denotes bankruptcy for 1, nonbankruptcy for 0

## Descriptive Statistics for Nonbankrupt Firms

N Obs	Minimum	Maximum	Mean	Std Dev
176	-11.2225086	4.1124162	-3.6243104	2.1193865

## Descriptive Statistics for Bankrupt Firms

N Obs	Minimum	Maximum	Mean	Std Dev
88	-2.4320484	17.6668115	3.6237575	3.5765666

## APPENDIX VI

### Conditional Probabilities of Bankruptcy for 264 Companies and Descriptive Statistics by Group Using Logit Procedure

OBS	GROUP*	Conditional Probability	OBS	GROUP*	Conditional Probability
1	0	0.61531	51	0	0.99995
2	0	0.99983	52	0	0.97947
3	0	0.99999	53	0	0.99900
4	0	0.91823	54	0	0.93122
5	0	0.93805	55	0	0.99996
6	0	0.73833	56	0	0.99727
7	0	0.99342	57	0	0.99817
8	0	0.99992	58	0	0.99991
9	0	0.99995	59	0	0.99815
10	0	0.99547	60	0	0.99982
11	0	0.99216	61	0	0.99917
12	0	1.00000	62	0	1.00000
13	0	0.91382	63	0	0.89928
14	0	0.94788	64	0	0.99461
15	0	0.87694	65	0	0.99988
16	0	0.99510	66	0	0.99880
17	0	0.82565	67	0	0.99813
18	0	0.06839	68	0	0.90593
19	0	0.69650	69	0	0.99394
20	0	0.85773	70	0	0.99743
21	0	0.40620	71	0	0.80995
22	0	0.91472	72	0	0.99275
23	0	0.99789	73	0	0.99997
24	0	0.99746	74	0	0.99860
25	0	0.99414	75	0	0.97296
26	0	0.04824	76	0	0.99997
27	0	0.99901	77	0	0.71624
28	0	0.99420	78	0	0.99710
29	0	0.99905	79	0	0.40153
30	0	0.98806	80	0	0.78725
31	0	0.96169	81	0	0.94543
32	0	0.99998	82	0	0.99900
33	0	0.98325	83	0	0.99855
34	0	0.46924	84	0	1.00000
35	0	0.04760	85	0	0.27882
36	0	0.99869	86	0	0.99995
37	0	0.99979	87	0	0.38886
38	0	0.98039	88	0	0.99999
39	0	0.99998	89	1	0.99055
40	0	0.99820	90	1	0.00009
41	0	0.99998	91	1	0.08700
42	0	0.99800	92	1	0.00031
43	0	0.98371	93	1	0.00171
44	0	0.99820	94	1	0.01468
45	0	0.99906	95	1	0.00001
46	0	0.99762	96	1	0.12256
47	0	0.99999	97	1	0.00006
48	0	0.93991	98	1	0.00161
49	0	0.99999	99	1	0.01549
50	0	0.99995	100	1	0.01599

## APPENDIX VI

(continue)

### Conditional Probabilities of Bankruptcy for 264 Companies and Descriptive Statistics by Group Using Logit Procedure

OBS	GROUP*	Conditional Probability	OBS	GROUP*	Conditional Probability
101	1	0.00012	151	1	0.00006
102	1	0.00972	152	1	0.00012
103	1	0.00111	153	1	0.06365
104	1	0.00076	154	1	0.09920
105	1	0.51457	155	1	0.00611
106	1	0.00046	156	1	0.00012
107	1	0.05305	157	1	0.00003
108	1	0.03418	158	1	0.00864
109	1	0.08376	159	1	0.08593
110	1	0.00018	160	1	0.00007
111	1	0.00035	161	1	0.00143
112	1	0.00000	162	1	0.00102
113	1	0.00001	163	1	0.03152
114	1	0.00353	164	1	0.00002
115	1	0.03495	165	1	0.00007
116	1	0.00093	166	1	0.00121
117	1	0.00036	167	1	0.00687
118	1	0.00015	168	1	0.00185
119	1	0.08567	169	1	0.00041
120	1	0.01886	170	1	0.00067
121	1	0.00118	171	1	0.50427
122	1	0.01842	172	1	0.08032
123	1	0.03509	173	1	0.00003
124	1	0.00688	174	1	0.20164
125	1	0.00020	175	1	0.01197
126	1	0.00010	176	1	0.13275
127	1	0.00003	177	1	0.00911
128	1	0.11761	178	1	0.00356
129	1	0.066571	179	1	0.00170
130	1	0.005755	180	1	0.00016
131	1	0.000773	181	1	0.00005
132	1	0.017259	182	1	0.00709
133	1	0.013237	183	1	0.43331
134	1	0.000156	184	1	0.10043
135	1	0.000026	185	1	0.53355
136	1	0.019662	186	1	0.07305
137	1	0.041955	187	1	0.00062
138	1	0.003666	188	1	0.00090
139	1	0.000006	189	1	0.00012
140	1	0.000070	190	1	0.00865
141	1	0.000858	191	1	0.03907
142	1	0.000205	192	1	0.00070
143	1	0.016677	193	1	0.01015
144	1	0.000002	194	1	0.05371
145	1	0.01268	195	1	0.03612
146	1	0.00208	196	1	0.10072
147	1	0.10471	197	1	0.00087
148	1	0.00243	198	1	0.00683
149	1	0.09548	199	1	0.00052
150	1	0.01164	200	1	0.01580

## APPENDIX VI

(continue)

### Conditional Probabilities of Bankruptcy for 264 Companies and Descriptive Statistics by Group Using Logit Procedure

OBS	GROUP*	Conditional Probability	OBS	GROUP*	Conditional Probability
201	1	0.07456	233	1	0.00001
202	1	0.00034	234	1	0.01862
203	1	0.07774	235	1	0.00308
204	1	0.00001	236	1	0.00226
205	1	0.00127	237	1	0.00732
206	1	0.00007	238	1	0.00010
207	1	0.00194	239	1	0.00165
208	1	0.05160	240	1	0.00011
209	1	0.00000	241	1	0.00036
210	1	0.00031	242	1	0.00957
211	1	0.00002	243	1	0.00098
212	1	0.00051	244	1	0.25028
213	1	0.28167	245	1	0.03269
214	1	0.98607	246	1	0.01147
215	1	0.63125	247	1	0.00274
216	1	0.00884	248	1	0.00149
217	1	0.00001	249	1	0.00004
218	1	0.03830	250	1	0.00001
219	1	0.01823	251	1	0.00003
220	1	0.00011	252	1	0.00455
221	1	0.00268	253	1	0.00267
222	1	0.36061	254	1	0.00007
223	1	0.06479	255	1	0.00024
224	1	0.00422	256	1	0.64302
225	1	0.00442	257	1	0.01480
226	1	0.00327	258	1	0.00011
227	1	0.00899	259	1	0.07555
228	1	0.00007	260	1	0.02705
229	1	0.01425	261	1	0.15604
230	1	0.10735	262	1	0.49069
231	1	0.07598	263	1	0.00302
232	1	0.00090	264	1	0.00060

\* denotes bankruptcy for 0, nonbankruptcy for 1

#### Descriptive Statistics for Bankrupt Firms

NObs	N	Minimum	Maximum	Mean	Std Dev
88	88	0.0476000	1.000000	0.8981190	0.2210777

#### Descriptive Statistics for Nonbankrupt Firms

NObs	N	Minimum	Maximum	Mean	Std Dev
176	176	0	0.9860700	0.0508874	0.1332954

## APPENDIX VII

### The Intermediate Values Z of 264 Companies and Descriptive Statistics by Group Using Logit Procedure

OBS	GROUP*	Intermediate Value Z	OBS	GROUP*	Intermediate Value Z
1	0	0.4697	51	0	3.8649
2	0	8.6645	52	0	6.9036
3	0	11.5346	53	0	2.6056
4	0	2.4186	54	0	10.1237
5	0	2.7175	55	0	5.9009
6	0	1.0373	56	0	6.3014
7	0	5.0168	57	0	9.2686
8	0	9.3927	58	0	6.2923
9	0	10.0037	59	0	8.6150
10	0	5.3927	60	0	7.0942
11	0	4.8411	61	0	13.3225
12	0	12.7599	62	0	2.1893
13	0	2.3612	63	0	5.2172
14	0	2.9006	64	0	9.0478
15	0	1.9638	65	0	6.7254
16	0	5.3137	66	0	6.2791
17	0	1.5551	67	0	2.2649
18	0	-2.6117	68	0	5.1000
19	0	0.8307	69	0	5.9627
20	0	1.7966	70	0	1.4497
21	0	-0.3797	71	0	4.9195
22	0	2.3727	72	0	10.3519
23	0	6.1600	73	0	6.5734
24	0	5.9748	74	0	3.5831
25	0	5.1329	75	0	10.5295
26	0	-2.9821	76	0	0.9259
27	0	6.9215	77	0	5.8394
28	0	5.1441	78	0	-0.3991
29	0	6.9529	79	0	1.3084
30	0	4.4161	80	0	2.8522
31	0	3.2229	81	0	6.9054
32	0	10.8049	82	0	6.5317
33	0	4.0725	83	0	15.8825
34	0	-0.1232	84	0	-0.9503
35	0	-2.9961	85	0	9.8981
36	0	6.6384	86	0	-0.4521
37	0	8.4695	87	0	11.1174
38	0	3.9117	88	0	4.6520
39	0	11.0531	89	1	-9.3405
40	0	6.3156	90	1	-2.3508
41	0	10.7166	91	1	-8.0654
42	0	6.2110	92	1	-6.3700
43	0	4.1008	93	1	-4.2062
44	0	6.3157	94	1	-12.1159
45	0	6.9698	95	1	-1.9684
46	0	6.0391	96	1	-9.6701
47	0	11.8760	97	1	-6.4289
48	0	2.7500	98	1	-4.1522
49	0	11.5521	99	1	-4.1197
50	0	9.8754	100	1	-8.6250

## APPENDIX VII

(continue)

### The Intermediate Values Z of 264 Companies and Descriptive Statistics by Group Using Logit Procedure

OBS	GROUP*	Intermediate value Z	OBS	GROUP*	Intermediate Value Z
101	1	-8.9961	151	1	-9.7794
102	1	-4.6238	152	1	-8.9952
103	1	-6.8021	153	1	-2.6886
104	1	-7.1830	154	1	-2.2061
105	1	0.0583	155	1	-5.0919
106	1	-7.6737	156	1	-9.0695
107	1	-2.8820	157	1	-10.2870
108	1	-3.3413	158	1	-4.7425
109	1	-2.3923	159	1	-2.3644
110	1	-8.6446	160	1	-9.5682
111	1	-7.9620	161	1	-6.5466
112	1	-14.1105	162	1	-6.8864
113	1	-11.9374	163	1	-3.4252
114	1	-5.6421	164	1	-10.6353
115	1	-3.3182	165	1	-9.6002
116	1	-6.9745	166	1	-6.7143
117	1	-7.9340	167	1	-4.9735
118	1	-8.8198	168	1	-6.2926
119	1	-2.3677	169	1	-7.7884
120	1	-3.9515	170	1	-7.3150
121	1	-6.7452	171	1	0.0171
122	1	-3.9760	172	1	-2.4380
123	1	-3.3141	173	1	-10.4971
124	1	-4.9723	174	1	-1.3761
125	1	-8.5093	175	1	-4.4136
126	1	-9.2368	176	1	-1.8769
127	1	-10.5958	177	1	-4.6895
128	1	-2.0153	178	1	-5.6340
129	1	-2.6406	179	1	-6.3741
130	1	-5.1519	180	1	-8.7458
131	1	-7.1647	181	1	-9.9370
132	1	-4.0420	182	1	-4.9414
133	1	-4.3114	183	1	-0.2684
134	1	-8.7684	184	1	-2.1925
135	1	-10.5438	185	1	0.1344
136	1	-3.9092	186	1	-2.5408
137	1	-3.1283	187	1	-7.3916
138	1	-5.6050	188	1	-7.0092
139	1	-11.9478	189	1	-9.0047
140	1	-9.5661	190	1	-4.7414
141	1	-7.0602	191	1	-3.2026
142	1	-8.4916	192	1	-7.2569
143	1	-4.0769	193	1	-4.5798
144	1	-12.9688	194	1	-2.8689
145	1	-4.3550	195	1	-3.2840
146	1	-6.1744	196	1	-2.1893
147	1	-2.1459	197	1	-7.0430
148	1	-6.0189	198	1	-4.9792
149	1	-2.2485	199	1	-7.5686
150	1	-4.4418	200	1	-4.1318

## APPENDIX VII

(continue)

### The Intermediate Values Z of 264 Companies and Descriptive Statistics by Group Using Logit Procedure

OBS	GROUP*	Intermediate Value Z	OBS	GROUP*	Intermediate Value Z
201	1	-2.5187	233	1	-11.4563
202	1	-7.9766	234	1	-3.9645
203	1	-2.4735	235	1	-5.7813
204	1	-11.2878	236	1	-6.0923
205	1	-6.6650	237	1	-4.9095
206	1	-9.5406	238	1	-9.2101
207	1	-6.2417	239	1	-6.4063
208	1	-2.9112	240	1	-9.1217
209	1	-12.3445	241	1	-7.9154
210	1	-8.0849	242	1	-4.6398
211	1	-11.0721	243	1	-6.9225
212	1	-7.5832	244	1	-1.0971
213	1	-0.9362	245	1	-3.3873
214	1	4.2594	246	1	-4.4562
215	1	0.5376	247	1	-5.8953
216	1	-4.7197	248	1	-6.5097
217	1	-11.2448	249	1	-10.1182
218	1	-3.2232	250	1	-11.7757
219	1	-3.9865	251	1	-10.5461
220	1	-9.1406	252	1	-5.3875
221	1	-5.9201	253	1	-5.9239
222	1	-0.5727	254	1	-9.6106
223	1	-2.6697	255	1	-8.3543
224	1	-5.4630	256	1	0.5885
225	1	-5.4167	257	1	-4.19805
226	1	-5.7185	258	1	-9.14287
227	1	-4.7022	259	1	-2.50436
228	1	-9.5385	260	1	-3.58263
229	1	-4.2364	261	1	-1.68802
230	1	-2.1181	262	1	-0.03723
231	1	-2.4982	263	1	-5.79915
232	1	-7.0067	264	1	-7.42005

\* denotes bankruptcy for 0, nonbankruptcy for 1

#### Descriptive Statistics for Bankrupt Firms

NObs	N	Minimum	Maximum	Mean	Std Dev
88	88	-2.9961000	15.8825000	5.4596114	3.9195392

#### Descriptive Statistics for Nonbankrupt Firms

NObs	N	Minimum	Maximum	Mean	Std Dev
176	176	-14.1105000	4.2594000	-5.8483647	3.2433033



## APPENDIX VIII

### The Predicted Values vs. Actual Values of 264 Companies and Descriptive Statistics by Group Using GDR Neural Network Approach

OBS	GROUP*	Conditional Probability	OBS	GROUP*	Conditional Probability
1	1.000000	0.954131	49	1.000000	0.995116
2	1.000000	0.954131	50	1.000000	0.995116
3	1.000000	0.990879	51	1.000000	0.990879
4	1.000000	0.954131	52	1.000000	0.954131
5	1.000000	0.954131	53	1.000000	0.954131
6	1.000000	0.954131	54	1.000000	0.995116
7	1.000000	0.990879	55	1.000000	0.954131
8	1.000000	0.995116	56	1.000000	0.954131
9	1.000000	0.990879	57	1.000000	0.990879
10	1.000000	0.995116	58	1.000000	0.954131
11	1.000000	0.990879	59	1.000000	0.995116
12	1.000000	0.990879	60	1.000000	0.995116
13	1.000000	0.954131	61	1.000000	0.990879
14	1.000000	0.954131	62	1.000000	0.954131
15	1.000000	0.954131	63	1.000000	0.954131
16	1.000000	0.954131	64	1.000000	0.990879
17	1.000000	0.954131	65	1.000000	0.954131
18	1.000000	0.019521	66	1.000000	0.954131
19	1.000000	0.954131	67	1.000000	0.954131
20	1.000000	0.954131	68	1.000000	0.990894
21	1.000000	0.954131	69	1.000000	0.954131
22	1.000000	0.954131	70	1.000000	0.954131
23	1.000000	0.954131	71	1.000000	0.954131
24	1.000000	0.954131	72	1.000000	0.990879
25	1.000000	0.954131	73	1.000000	0.995116
26	1.000000	0.154733	74	1.000000	0.954131
27	1.000000	0.990879	75	1.000000	0.990879
28	1.000000	0.954131	76	1.000000	0.954113
29	1.000000	0.990879	77	1.000000	0.995116
30	1.000000	0.954131	78	1.000000	0.333938
31	1.000000	0.954131	79	1.000000	0.954131
32	1.000000	0.990879	80	1.000000	0.954131
33	1.000000	0.954231	81	1.000000	0.954131
34	1.000000	0.333938	82	1.000000	0.954131
35	1.000000	0.056816	83	1.000000	0.995116
36	1.000000	0.954131	84	1.000000	0.333938
37	1.000000	0.990879	85	1.000000	0.990879
38	1.000000	0.954131	86	1.000000	0.333938
39	1.000000	0.995116	87	1.000000	0.990879
40	1.000000	0.990879	88	1.000000	0.954131
41	1.000000	0.995116	89	0.000000	0.004883
42	1.000000	0.954131	90	0.000000	0.129423
43	1.000000	0.954131	91	0.000000	0.004883
44	1.000000	0.990879	92	0.000000	0.004913
45	1.000000	0.990879	93	0.000000	0.039550
46	1.000000	0.990879	94	0.000000	0.004883
47	1.000000	0.995116	95	0.000000	0.129423
48	1.000000	0.954131	96	0.000000	0.004883

## APPENDIX VIII

(continue)

### The Predicted Values vs. Actual Values of 264 Companies and Descriptive Statistics by Group Using GDR Neural Network Approach

OBS	GROUP*	Conditional Probability	OBS	GROUP*	Conditional Probability
97	0.000000	0.004884	150	0.000000	0.007719
98	0.000000	0.011518	151	0.000000	0.004883
99	0.000000	0.011290	152	0.000000	0.004883
100	0.000000	0.004883	153	0.000000	0.019522
101	0.000000	0.004883	154	0.000000	0.101184
102	0.000000	0.006907	155	0.000000	0.005118
103	0.000000	0.004883	156	0.000000	0.004883
104	0.000000	0.004883	157	0.000000	0.004883
105	0.000000	0.083746	158	0.000000	0.006918
106	0.000000	0.004883	159	0.000000	0.073458
107	0.000000	0.084857	160	0.000000	0.004883
108	0.000000	0.032272	161	0.000000	0.004883
109	0.000000	0.129423	162	0.000000	0.004883
110	0.000000	0.004883	163	0.000000	0.012558
111	0.000000	0.004883	164	0.000000	0.004883
112	0.000000	0.004883	165	0.000000	0.004883
113	0.000000	0.004883	166	0.000000	0.004884
114	0.000000	0.005002	167	0.000000	0.005848
115	0.000000	0.019420	168	0.000000	0.004884
116	0.000000	0.004883	169	0.000000	0.004883
117	0.000000	0.004883	170	0.000000	0.004883
118	0.000000	0.004883	171	0.000000	0.129423
119	0.000000	0.129423	172	0.000000	0.019521
120	0.000000	0.012324	173	0.000000	0.004883
121	0.000000	0.004886	174	0.000000	0.129423
122	0.000000	0.007497	175	0.000000	0.005691
123	0.000000	0.129421	176	0.000000	0.086864
124	0.000000	0.004994	177	0.000000	0.005863
125	0.000000	0.004883	178	0.000000	0.004883
126	0.000000	0.004883	179	0.000000	0.004883
127	0.000000	0.004883	180	0.000000	0.004883
128	0.000000	0.129423	181	0.000000	0.004883
129	0.000000	0.101184	182	0.000000	0.005599
130	0.000000	0.004883	183	0.000000	0.333938
131	0.000000	0.004883	184	0.000000	0.129423
132	0.000000	0.007749	185	0.000000	0.333938
133	0.000000	0.007537	186	0.000000	0.034102
134	0.000000	0.004883	187	0.000000	0.004883
135	0.000000	0.004883	188	0.000000	0.004883
136	0.000000	0.017085	189	0.000000	0.004883
137	0.000000	0.019129	190	0.000000	0.006237
138	0.000000	0.004988	191	0.000000	0.034140
139	0.000000	0.004883	192	0.000000	0.004883
140	0.000000	0.004883	193	0.000000	0.006231
141	0.000000	0.004883	194	0.000000	0.019237
142	0.000000	0.004883	195	0.000000	0.101008
143	0.000000	0.017878	196	0.000000	0.019518
144	0.000000	0.004883	197	0.000000	0.004884
145	0.000000	0.006841	198	0.000000	0.005606
146	0.000000	0.004884	199	0.000000	0.004883
147	0.000000	0.129423	200	0.000000	0.011223
148	0.000000	0.004892	201	0.000000	0.129422
149	0.000000	0.101184	202	0.000000	0.004883

## APPENDIX VIII

(continue)

### The Predicted Values vs. Actual Values of 264 Companies and Descriptive Statistics by Group Using GDR Neural Network Approach

OBS	GROUP*	Conditional Probability	OBS	GROUP*	Conditional Probability
203	0.000000	0.129423	234	0.000000	0.019073
204	0.000000	0.004883	235	0.000000	0.004883
205	0.000000	0.004883	236	0.000000	0.004891
206	0.000000	0.004883	237	0.000000	0.007490
207	0.000000	0.004905	238	0.000000	0.004883
208	0.000000	0.056811	239	0.000000	0.005203
209	0.000000	0.004883	240	0.000000	0.004883
210	0.000000	0.004883	241	0.000000	0.004883
211	0.000000	0.004883	242	0.000000	0.005832
212	0.000000	0.004883	243	0.000000	0.004883
213	0.000000	0.333938	244	0.000000	0.333938
214	0.000000	0.954131	245	0.000000	0.056366
215	0.000000	0.333938	246	0.000000	0.004883
216	0.000000	0.005565	247	0.000000	0.004923
217	0.000000	0.004883	248	0.000000	0.004907
218	0.000000	0.056816	249	0.000000	0.004883
219	0.000000	0.011378	250	0.000000	0.004883
220	0.000000	0.004883	251	0.000000	0.004883
221	0.000000	0.004883	252	0.000000	0.006067
222	0.000000	0.333933	253	0.000000	0.004883
223	0.000000	0.073458	254	0.000000	0.004883
224	0.000000	0.004912	255	0.000000	0.004883
225	0.000000	0.005320	256	0.000000	0.954131
226	0.000000	0.005134	257	0.000000	0.004883
227	0.000000	0.005042	258	0.000000	0.004883
228	0.000000	0.004883	259	0.000000	0.007390
229	0.000000	0.006492	260	0.000000	0.019520
230	0.000000	0.129423	261	0.000000	0.129423
231	0.000000	0.129423	262	0.000000	0.333938
232	0.000000	0.004883	263	0.000000	0.006068
233	0.000000	0.004883	264	0.000000	0.004883

\* denotes bankruptcy for 1, nonbankruptcy for 0

#### Descriptive Statistics for Nonbankrupt Firms

NObs	N	Minimum	Maximum	Mean	Std Dev
176	176	0.0048830	0.9541310	0.0465145	0.1211258

#### Descriptive Statistics for Bankrupt Firms

NObs	N	Minimum	Maximum	Mean	Std Dev
88	88	0.0195210	0.9951160	0.9112806	0.2075093

## APPENDIX IX

### The Intermediate Values Z of 264 Companies and Descriptive Statistics by Group Using GDR Neural Network Approach

OBS	GROUP*	Intermediate Value Z	OBS	GROUP*	Intermediate Value Z
1	1.000000	3.03501	49	1.000000	5.31689
2	1.000000	3.03501	50	1.000000	5.31689
3	1.000000	4.68801	51	1.000000	4.68801
4	1.000000	3.03501	52	1.000000	3.03501
5	1.000000	3.03501	53	1.000000	3.03501
6	1.000000	3.03501	54	1.000000	5.31689
7	1.000000	4.68801	55	1.000000	3.03501
8	1.000000	5.31689	56	1.000000	3.03501
9	1.000000	4.68801	57	1.000000	4.68801
10	1.000000	5.31689	58	1.000000	3.03501
11	1.000000	4.68801	59	1.000000	5.31689
12	1.000000	4.68801	60	1.000000	5.31689
13	1.000000	3.03501	61	1.000000	4.68801
14	1.000000	3.03501	62	1.000000	3.03501
15	1.000000	3.03501	63	1.000000	3.03501
16	1.000000	3.03501	64	1.000000	4.68801
17	1.000000	3.03501	65	1.000000	3.03501
18	1.000000	-3.91655	66	1.000000	3.03501
19	1.000000	3.03501	67	1.000000	3.03501
20	1.000000	3.03501	68	1.000000	4.68967
21	1.000000	3.03501	69	1.000000	3.03501
22	1.000000	3.03501	70	1.000000	3.03501
23	1.000000	3.03501	71	1.000000	3.03501
24	1.000000	3.03501	72	1.000000	4.68801
25	1.000000	3.03501	73	1.000000	5.31689
26	1.000000	-1.69795	74	1.000000	3.03501
27	1.000000	4.68801	75	1.000000	4.68801
28	1.000000	3.03501	76	1.000000	3.03460
29	1.000000	4.68801	77	1.000000	5.31689
30	1.000000	3.03501	78	1.000000	-0.69043
31	1.000000	3.03501	79	1.000000	3.03501
32	1.000000	4.68801	80	1.000000	3.03501
33	1.000000	3.03730	81	1.000000	3.03501
34	1.000000	-0.69043	82	1.000000	3.03501
35	1.000000	-2.80944	83	1.000000	5.31689
36	1.000000	3.03501	84	1.000000	-0.69043
37	1.000000	4.68801	85	1.000000	4.68801
38	1.000000	3.03501	86	1.000000	-0.69043
39	1.000000	5.31689	87	1.000000	4.68801
40	1.000000	4.68801	88	1.000000	3.03501
41	1.000000	5.31689	89	0.000000	-5.31710
42	1.000000	3.03501	90	0.000000	-1.90607
43	1.000000	3.03501	91	0.000000	-5.31710
44	1.000000	4.68801	92	0.000000	-5.31095
45	1.000000	4.68801	93	0.000000	-3.18984
46	1.000000	4.68801	94	0.000000	-5.31710
47	1.000000	5.31689	95	0.000000	-1.90607
48	1.000000	3.03501	96	0.000000	-5.31710

## APPENDIX IX

(continue)

### The Intermediate Values Z of 264 Companies and Descriptive Statistics by Group Using GDR Neural Network Approach

OBS	GROUP*	Intermediate Value Z	OBS	GROUP*	Intermediate Value Z
97	0.000000	-5.31689	150	0.00000	-4.85632
98	0.000000	-4.45226	151	0.00000	-5.31710
99	0.000000	-4.47248	152	0.00000	-5.31710
100	0.000000	-5.31710	153	0.00000	-3.91650
101	0.000000	-5.31710	154	0.00000	-2.18414
102	0.000000	-4.96829	155	0.00000	-5.26986
103	0.000000	-5.31710	156	0.00000	-5.31710
104	0.000000	-5.31710	157	0.00000	-5.31710
105	0.000000	-2.39251	158	0.00000	-4.96669
106	0.000000	-5.31710	159	0.00000	-2.53475
107	0.000000	-2.37811	160	0.00000	-5.31710
108	0.000000	-3.40075	161	0.00000	-5.31710
109	0.000000	-1.90607	162	0.00000	-5.31710
110	0.000000	-5.31710	163	0.00000	-4.36476
111	0.000000	-5.31710	164	0.00000	-5.31710
112	0.000000	-5.31710	165	0.00000	-5.31710
113	0.000000	-5.31710	166	0.00000	-5.31689
114	0.000000	-5.29290	167	0.00000	-5.13579
115	0.000000	-3.92184	168	0.00000	-5.31689
116	0.000000	-5.31710	169	0.00000	-5.31710
117	0.000000	-5.31710	170	0.00000	-5.31710
118	0.000000	-5.31710	171	0.00000	-1.90607
119	0.000000	-1.90607	172	0.00000	-3.91655
120	0.000000	-4.38381	173	0.00000	-5.31710
121	0.000000	-5.31648	174	0.00000	-1.90607
122	0.000000	-4.88573	175	0.00000	-5.16316
123	0.000000	-1.90609	176	0.00000	-2.35254
124	0.000000	-5.29451	177	0.00000	-5.13321
125	0.000000	-5.31710	178	0.00000	-5.31710
126	0.000000	-5.31710	179	0.00000	-5.31710
127	0.000000	-5.31710	180	0.00000	-5.31710
128	0.000000	-1.90607	181	0.00000	-5.31710
129	0.000000	-2.18414	182	0.00000	-5.17955
130	0.000000	-5.31710	183	0.00000	-0.69043
131	0.000000	-5.31710	184	0.00000	-1.90607
132	0.000000	-4.85241	185	0.00000	-0.69043
133	0.000000	-4.88037	186	0.00000	-3.34370
134	0.000000	-5.31710	187	0.00000	-5.31710
135	0.000000	-5.31710	188	0.00000	-5.31710
136	0.000000	-4.05232	189	0.00000	-5.31710
137	0.000000	-3.93724	190	0.00000	-5.07100
138	0.000000	-5.29572	191	0.00000	-3.34255
139	0.000000	-5.31710	192	0.00000	-5.31710
140	0.000000	-5.31710	193	0.00000	-5.07197
141	0.000000	-5.31710	194	0.00000	-3.93150
142	0.000000	-5.31710	195	0.00000	-2.18607
143	0.000000	-4.00614	196	0.00000	-3.91671
144	0.000000	-5.31710	197	0.00000	-5.31689
145	0.000000	-4.97796	198	0.00000	-5.17830
146	0.000000	-5.31689	199	0.00000	-5.31710
147	0.000000	-1.90607	200	0.00000	-4.47850
148	0.000000	-5.31525	201	0.00000	-1.90608
149	0.000000	-2.18414	202	0.00000	-5.31710

## APPENDIX IX

(continue)

### The Intermediate Values Z of 264 Companies and Descriptive Statistics by Group Using GDR Neural Network Approach

OBS	GROUP*	Intermediate Value Z	OBS	GROUP*	Intermediate Value Z
203	0.000000	-1.90607	234	0.000000	-3.94022
204	0.000000	-5.31710	235	0.000000	-5.31710
205	0.000000	-5.31710	236	0.000000	-5.31546
206	0.000000	-5.31710	237	0.000000	-4.88667
207	0.000000	-5.31258	238	0.000000	-5.31710
208	0.000000	-2.80954	239	0.000000	-5.25330
209	0.000000	-5.31710	240	0.000000	-5.31710
210	0.000000	-5.31710	241	0.000000	-5.31710
211	0.000000	-5.31710	242	0.000000	-5.13855
212	0.000000	-5.31710	243	0.000000	-5.31710
213	0.000000	-0.69043	244	0.000000	-0.69043
214	0.000000	3.03501	245	0.000000	-2.81787
215	0.000000	-0.69043	246	0.000000	-5.31710
216	0.000000	-5.18568	247	0.000000	-5.30890
217	0.000000	-5.31710	248	0.000000	-5.31217
218	0.000000	-2.80944	249	0.000000	-5.31710
219	0.000000	-4.46463	250	0.000000	-5.31710
220	0.000000	-5.31710	251	0.000000	-5.31710
221	0.000000	-5.31710	252	0.000000	-5.09881
222	0.000000	-0.69045	253	0.000000	-5.31710
223	0.000000	-2.53475	254	0.000000	-5.31710
224	0.000000	-5.31115	255	0.000000	-5.31710
225	0.000000	-5.23095	256	0.000000	3.03501
226	0.000000	-5.26672	257	0.000000	-5.31710
227	0.000000	-5.28490	258	0.000000	-5.31710
228	0.000000	-5.31710	259	0.000000	-4.90021
229	0.000000	-5.03067	260	0.000000	-3.91660
230	0.000000	-1.90607	261	0.000000	-1.90607
231	0.000000	-1.90607	262	0.000000	-0.69043
232	0.000000	-5.31710	263	0.000000	-5.09864
233	0.000000	-5.31710	264	0.000000	-5.31710

\* denotes bankruptcy for 1, nonbankruptcy for 0

#### Descriptive Statistics for Nonbankrupt Firms

NObs	N	Minimum	Maximum	Mean	Std Dev
176	176	-5.3171000	3.0350100	-4.3621649	1.5936182

#### Descriptive Statistics for Bankrupt Firms

NObs	N	Minimum	Maximum	Mean	Std Dev
88	88	-3.9165500	5.3168900	3.4168651	1.7791123

## APPENDIX X

### The Predicted Values vs. Actual Values of 264 Companies and Descriptive Statistics by Groups Using Projection Neural Network Approach

OBS	GROUP*	Conditional Probability	OBS	GROUP*	Conditional Probability
1	1.000000	0.896663	49	1.000000	0.980614
2	1.000000	0.977718	50	1.000000	0.981770
3	1.000000	0.978727	51	1.000000	0.980633
4	1.000000	0.592849	52	1.000000	0.974152
5	1.000000	0.893021	53	1.000000	0.731432
6	1.000000	0.877478	54	1.000000	0.980126
7	1.000000	0.978226	55	1.000000	0.946164
8	1.000000	0.978567	56	1.000000	0.974930
9	1.000000	0.980060	57	1.000000	0.979607
10	1.000000	0.980422	58	1.000000	0.976726
11	1.000000	0.979380	59	1.000000	0.980876
12	1.000000	0.981356	60	1.000000	0.979915
13	1.000000	0.810727	61	1.000000	0.980904
14	1.000000	0.521412	62	1.000000	0.829099
15	1.000000	0.786870	63	1.000000	0.861232
16	1.000000	0.965272	64	1.000000	0.977679
17	1.000000	0.703180	65	1.000000	0.977508
18	1.000000	0.022325	66	1.000000	0.974578
19	1.000000	0.623915	67	1.000000	0.874387
20	1.000000	0.730932	68	1.000000	0.981482
21	1.000000	0.586246	69	1.000000	0.973846
22	1.000000	0.718382	70	1.000000	0.505306
23	1.000000	0.970255	71	1.000000	0.966702
24	1.000000	0.976433	72	1.000000	0.980943
25	1.000000	0.969679	73	1.000000	0.980041
26	1.000000	0.050355	74	1.000000	0.797824
27	1.000000	0.979918	75	1.000000	0.981003
28	1.000000	0.968991	76	1.000000	0.753441
29	1.000000	0.979293	77	1.000000	0.976423
30	1.000000	0.972212	78	1.000000	0.385868
31	1.000000	0.917724	79	1.000000	0.935732
32	1.000000	0.980102	80	1.000000	0.867604
33	1.000000	0.955258	81	1.000000	0.973696
34	1.000000	0.219125	82	1.000000	0.972343
35	1.000000	0.033031	83	1.000000	0.981630
36	1.000000	0.976649	84	1.000000	0.172142
37	1.000000	0.979638	85	1.000000	0.979681
38	1.000000	0.948012	86	1.000000	0.151102
39	1.000000	0.981621	87	1.000000	0.978229
40	1.000000	0.979706	88	1.000000	0.950395
41	1.000000	0.966995	89	0.000000	0.002542
42	1.000000	0.975277	90	0.000000	0.106064
43	1.000000	0.947228	91	0.000000	0.002862
44	1.000000	0.979176	92	0.000000	0.006512
45	1.000000	0.978377	93	0.000000	0.041051
46	1.000000	0.975741	94	0.000000	0.002191
47	1.000000	0.981706	95	0.000000	0.080259
48	1.000000	0.959832	96	0.000000	0.002255

## APPENDIX X

(continue)

### The Predicted Values vs. Actual Values of 264 Companies and Descriptive Statistics by Groups Using Projection Neural Network Approach

OBS	GROUP*	Conditional Probability	OBS	GROUP*	Conditional Probability
97	0.000000	0.005240	150	0.000000	0.024963
98	0.000000	0.036979	151	0.000000	0.002274
99	0.000000	0.029741	152	0.000000	0.002302
100	0.000000	0.006394	153	0.000000	0.038972
101	0.000000	0.002540	154	0.000000	0.045787
102	0.000000	0.027313	155	0.000000	0.003272
103	0.000000	0.002724	156	0.000000	0.002435
104	0.000000	0.002836	157	0.000000	0.002380
105	0.000000	0.114591	158	0.000000	0.005227
106	0.000000	0.002892	159	0.000000	0.051087
107	0.000000	0.064587	160	0.000000	0.002222
108	0.000000	0.037180	161	0.000000	0.002269
109	0.000000	0.064281	162	0.000000	0.002531
110	0.000000	0.002573	163	0.000000	0.008651
111	0.000000	0.002517	164	0.000000	0.002194
112	0.000000	0.002260	165	0.000000	0.002382
113	0.000000	0.002287	166	0.000000	0.002418
114	0.000000	0.003121	167	0.000000	0.005606
115	0.000000	0.030906	168	0.000000	0.003750
116	0.000000	0.002395	169	0.000000	0.002365
117	0.000000	0.002517	170	0.000000	0.002981
118	0.000000	0.002506	171	0.000000	0.084243
119	0.000000	0.082172	172	0.000000	0.015133
120	0.000000	0.008126	173	0.000000	0.002324
121	0.000000	0.002503	174	0.000000	0.079857
122	0.000000	0.015621	175	0.000000	0.003188
123	0.000000	0.071130	176	0.000000	0.087483
124	0.000000	0.007535	177	0.000000	0.003654
125	0.000000	0.002470	178	0.000000	0.002950
126	0.000000	0.002271	179	0.000000	0.002864
127	0.000000	0.002301	180	0.000000	0.002268
128	0.000000	0.107440	181	0.000000	0.002337
129	0.000000	0.054631	182	0.000000	0.005290
130	0.000000	0.002308	183	0.000000	0.229697
131	0.000000	0.002657	184	0.000000	0.072135
132	0.000000	0.007957	185	0.000000	0.396657
133	0.000000	0.024691	186	0.000000	0.017025
134	0.000000	0.002291	187	0.000000	0.002464
135	0.000000	0.002186	188	0.000000	0.002902
136	0.000000	0.014075	189	0.000000	0.002281
137	0.000000	0.007594	190	0.000000	0.003555
138	0.000000	0.006060	191	0.000000	0.021202
139	0.000000	0.002179	192	0.000000	0.002712
140	0.000000	0.002364	193	0.000000	0.016361
141	0.000000	0.002978	194	0.000000	0.039864
142	0.000000	0.002622	195	0.000000	0.046071
143	0.000000	0.012568	196	0.000000	0.009517
144	0.000000	0.002092	197	0.000000	0.007220
145	0.000000	0.019092	198	0.000000	0.004885
146	0.000000	0.002620	199	0.000000	0.002359
147	0.000000	0.079863	200	0.000000	0.017225
148	0.000000	0.002888	201	0.000000	0.058561
149	0.000000	0.050249	202	0.000000	0.002259



## APPENDIX X

(continue)

### The Predicted Values vs. Actual Values of 264 Companies and Descriptive Statistics by Groups Using Projection Neural Network Approach

OBS	GROUP*	Conditional Probability	OBS	GROUP*	Conditional Probability
203	0.000000	0.101291	234	0.000000	0.020856
204	0.000000	0.002059	235	0.000000	0.002756
205	0.000000	0.002612	236	0.000000	0.003417
206	0.000000	0.002290	237	0.000000	0.023601
207	0.000000	0.003420	238	0.000000	0.003329
208	0.000000	0.045743	239	0.000000	0.004066
209	0.000000	0.002206	240	0.000000	0.002261
210	0.000000	0.002368	241	0.000000	0.003142
211	0.000000	0.002510	242	0.000000	0.006587
212	0.000000	0.003681	243	0.000000	0.007830
213	0.000000	0.312320	244	0.000000	0.152790
214	0.000000	0.977952	245	0.000000	0.038475
215	0.000000	0.249695	246	0.000000	0.002154
216	0.000000	0.009646	247	0.000000	0.002469
217	0.000000	0.002302	248	0.000000	0.003486
218	0.000000	0.034843	249	0.000000	0.002347
219	0.000000	0.025076	250	0.000000	0.002227
220	0.000000	0.002071	251	0.000000	0.002300
221	0.000000	0.002274	252	0.000000	0.022682
222	0.000000	0.138174	253	0.000000	0.002184
223	0.000000	0.036361	254	0.000000	0.002199
224	0.000000	0.005798	255	0.000000	0.002322
225	0.000000	0.008417	256	0.000000	0.465388
226	0.000000	0.022019	257	0.000000	0.002648
227	0.000000	0.004499	258	0.000000	0.002443
228	0.000000	0.002325	259	0.000000	0.268165
229	0.000000	0.004271	260	0.000000	0.030330
230	0.000000	0.077475	261	0.000000	0.075671
231	0.000000	0.093373	262	0.000000	0.227087
232	0.000000	0.007897	263	0.000000	0.003940
233	0.000000	0.002153	264	0.000000	0.002360

\* denotes bankruptcy for 1, nonbankruptcy for 0

#### Descriptive Statistics for Nonbankrupt Firms

N Obs	N	Minimum	Maximum	Mean	Std Dev
176	176	0.0020590	0.9779520	0.0360164	0.0970273

#### Descriptive Statistics for Bankrupt Firms

N Obs	N	Minimum	Maximum	Mean	Std Dev
88	88	0.0223250	0.9817700	0.8562935	0.2394382

## APPENDIX XI

### The Intermediate Values Z of 264 Companies and Descriptive Statistics by Group Using Projection Neural Network Approach

OBS	GROUP*	Intermediate Value Z	OBS	GROUP*	Intermediate Value Z
1	1.000000	2.16068	49	1.000000	3.92363
2	1.000000	3.78144	50	1.000000	3.98629
3	1.000000	3.82881	51	1.000000	3.92463
4	1.000000	0.37576	52	1.000000	3.62933
5	1.000000	2.12198	53	1.000000	1.00190
6	1.000000	1.96876	54	1.000000	3.89827
7	1.000000	3.80502	55	1.000000	2.86647
8	1.000000	3.82116	56	1.000000	3.66069
9	1.000000	3.89489	57	1.000000	3.87196
10	1.000000	3.91358	58	1.000000	3.73687
11	1.000000	3.86066	59	1.000000	3.93750
12	1.000000	3.96341	60	1.000000	3.88749
13	1.000000	1.45474	61	1.000000	3.93900
14	1.000000	0.08570	62	1.000000	1.57926
15	1.000000	1.30616	63	1.000000	1.82556
16	1.000000	3.32486	64	1.000000	3.77965
17	1.000000	0.86249	65	1.000000	3.77185
18	1.000000	-3.77947	66	1.000000	3.64639
19	1.000000	0.50620	67	1.000000	1.94032
20	1.000000	0.99936	68	1.000000	3.97032
21	1.000000	0.34847	69	1.000000	3.61725
22	1.000000	0.93645	70	1.000000	0.02122
23	1.000000	3.48490	71	1.000000	3.36839
24	1.000000	3.72406	72	1.000000	3.94108
25	1.000000	3.46512	73	1.000000	3.89391
26	1.000000	-2.93699	74	1.000000	1.37275
27	1.000000	3.88765	75	1.000000	3.94429
28	1.000000	3.44198	76	1.000000	1.11705
29	1.000000	3.85636	77	1.000000	3.72362
30	1.000000	3.55497	78	1.000000	-0.46471
31	1.000000	2.41182	79	1.000000	2.67827
32	1.000000	3.89704	80	1.000000	1.87994
33	1.000000	3.06107	81	1.000000	3.61138
34	1.000000	-1.27077	82	1.000000	3.55983
35	1.000000	-3.37672	83	1.000000	3.97850
36	1.000000	3.73349	84	1.000000	-1.57052
37	1.000000	3.87351	85	1.000000	3.87567
38	1.000000	2.90335	86	1.000000	-1.72598
39	1.000000	3.97800	87	1.000000	3.80516
40	1.000000	3.87693	88	1.000000	2.95279
41	1.000000	3.37753	89	0.000000	-5.97226
42	1.000000	3.67499	90	0.000000	-2.13159
43	1.000000	2.88756	91	0.000000	-5.85337
44	1.000000	3.85061	92	0.000000	-5.02758
45	1.000000	3.81214	93	0.000000	-3.15102
46	1.000000	3.69441	94	0.000000	-6.12120
47	1.000000	3.98272	95	0.000000	-2.43883
48	1.000000	3.17369	96	0.000000	-6.09235

# APPENDIX XI

(continue)

## The Intermediate Values Z of 264 Companies and Descriptive Statistics by Group Using Projection Neural Network Approach

OBS	GROUP*	Intermediate Value Z	OBS	GROUP*	Intermediate Value Z
97	0.000000	-5.24618	150	0.000000	-3.66508
98	0.000000	-3.25973	151	0.000000	-6.08394
99	0.000000	-3.48504	152	0.000000	-6.07167
100	0.000000	-5.04598	153	0.000000	-3.20516
101	0.000000	-5.97305	154	0.000000	-3.03689
102	0.000000	-3.57270	155	0.000000	-5.71908
103	0.000000	-5.90293	156	0.000000	-6.01537
104	0.000000	-5.86252	157	0.000000	-6.03827
105	0.000000	-2.04468	158	0.000000	-5.24868
106	0.000000	-5.84291	159	0.000000	-2.92179
107	0.000000	-2.67297	160	0.000000	-6.10712
108	0.000000	-3.25410	161	0.000000	-6.08614
109	0.000000	-2.67805	162	0.000000	-5.97661
110	0.000000	-5.96011	163	0.000000	-4.74139
111	0.000000	-5.98217	164	0.000000	-6.11983
112	0.000000	-6.09013	165	0.000000	-6.03743
113	0.000000	-6.07822	166	0.000000	-6.02239
114	0.000000	-5.76648	167	0.000000	-5.17830
115	0.000000	-3.44541	168	0.000000	-5.58224
116	0.000000	-6.03197	169	0.000000	-6.04461
117	0.000000	-5.98217	170	0.000000	-5.81251
118	0.000000	-5.98656	171	0.000000	-2.38605
119	0.000000	-2.41320	172	0.000000	-4.17563
120	0.000000	-4.80453	173	0.000000	-6.06214
121	0.000000	-5.98776	174	0.000000	-2.44429
122	0.000000	-4.14339	175	0.000000	-5.74517
123	0.000000	-2.56946	176	0.000000	-2.34476
124	0.000000	-4.88063	177	0.000000	-5.60827
125	0.000000	-6.00106	178	0.000000	-5.82300
126	0.000000	-6.08526	179	0.000000	-5.85267
127	0.000000	-6.07211	180	0.000000	-6.08659
128	0.000000	-2.11716	181	0.000000	-6.05655
129	0.000000	-2.85097	182	0.000000	-5.23663
130	0.000000	-6.06906	183	0.000000	-1.21002
131	0.000000	-5.92790	184	0.000000	-2.55435
132	0.000000	-4.82571	185	0.000000	-0.41941
133	0.000000	-3.67632	186	0.000000	-4.05590
134	0.000000	-6.07647	187	0.000000	-6.00350
135	0.000000	-6.12349	188	0.000000	-5.83945
136	0.000000	-4.24918	189	0.000000	-6.08086
137	0.000000	-4.87277	190	0.000000	-5.63584
138	0.000000	-5.09997	191	0.000000	-3.83223
139	0.000000	-6.12671	192	0.000000	-5.90735
140	0.000000	-6.04503	193	0.000000	-4.09636
141	0.000000	-5.81352	194	0.000000	-3.18160
142	0.000000	-5.94119	195	0.000000	-3.03041
143	0.000000	-4.36395	196	0.000000	-4.64511
144	0.000000	-6.16754	197	0.000000	-4.92365
145	0.000000	-3.93921	198	0.000000	-5.31669
146	0.000000	-5.94196	199	0.000000	-6.04716
147	0.000000	-2.44421	200	0.000000	-4.04402
148	0.000000	-5.84430	201	0.000000	-2.77734
149	0.000000	-2.93921	202	0.000000	-6.09057

## APPENDIX XI

(continue)

### The Intermediate Values Z of 264 Companies and Descriptive Statistics by Group Using Projection Neural Network Approach

OBS	GROUP*	Intermediate Value Z	OBS	GROUP*	Intermediate Value Z
203	0.000000	-2.18296	234	0.00000	-3.84904
204	0.000000	-6.18347	235	0.00000	-5.89122
205	0.000000	-5.94502	236	0.00000	-5.67557
206	0.000000	-6.07691	237	0.00000	-3.72258
207	0.000000	-5.67469	238	0.00000	-5.70175
208	0.000000	-3.03789	239	0.00000	-5.50102
209	0.000000	-6.11437	240	0.00000	-6.08968
210	0.000000	-6.04334	241	0.00000	-5.75975
211	0.000000	-5.98496	242	0.00000	-5.01605
212	0.000000	-5.60088	243	0.00000	-4.84193
213	0.000000	-0.78930	244	0.00000	-1.71288
214	0.000000	3.79224	245	0.00000	-3.21851
215	0.000000	-1.10024	246	0.00000	-6.13827
216	0.000000	-4.63152	247	0.00000	-6.00147
217	0.000000	-6.07167	248	0.00000	-5.65551
218	0.000000	-3.32144	249	0.00000	-6.05227
219	0.000000	-3.66045	250	0.00000	-6.10487
220	0.000000	-6.17765	251	0.00000	-6.07254
221	0.000000	-6.08394	252	0.00000	-3.76324
222	0.000000	-1.83054	253	0.00000	-6.12441
223	0.000000	-3.27722	254	0.00000	-6.11755
224	0.000000	-5.14443	255	0.00000	-6.06300
225	0.000000	-4.76905	256	0.00000	-0.13867
226	0.000000	-3.79358	257	0.00000	-5.93130
227	0.000000	-5.39939	258	0.00000	-6.01208
228	0.000000	-6.06171	259	0.00000	-1.00395
229	0.000000	-5.45163	260	0.00000	-3.46482
230	0.000000	-2.47716	261	0.00000	-2.50267
231	0.000000	-2.27313	262	0.00000	-1.22483
232	0.000000	-4.83334	263	0.00000	-5.53263
233	0.000000	-6.13874	264	0.00000	-6.04673

\* denotes bankruptcy for 1, nonbankruptcy for 0

#### Descriptive Statistics for Nonbankrupt Firms

NObs	N	Minimum	Maximum	Mean	Std Dev
176	176	-6.1834700	3.7922400	-4.7037655	1.6768469

#### Descriptive Statistics for Bankrupt Firms

NObs	N	Minimum	Maximum	Mean	Std Dev
88	88	-3.7794700	3.9862900	2.6354295	1.8319012